

Mujs, D. (2011) *Doing quantitative research in education with SPSS*. London: Sage.

4

Validity, reliability and generalisability

Three key concepts in quantitative methods are validity, reliability and generalisability. All three have got to do with *measurement*. Whenever we are doing quantitative research, we are trying to measure something. We might, for example, want to look at students' achievement in history. Achievement is a concept that we will have to try to measure, using a test, essay or portfolio. We might want to measure teachers' self-esteem. We would use a self-esteem instrument to do this. Measurement supplies the numbers we use in quantitative analyses. The question that follows from this is, how well are we measuring what we want to measure? If you wanted to measure your weight, you would want to be sure:

1. that you weren't measuring something else (such as height) instead
2. that whatever scale you were using wasn't completely erratic (e.g. gave you a different value every time you used it).

The same goes when we are trying to measure things in educational research. That is where *validity* (1) and *reliability* (2) come into play.

We saw in the previous chapter that we will usually take a sample, rather than studying the whole population. When we do this, what we really want to be able to do is to say something about the characteristics of the population rather than just our sample. This is called *generalising* from the sample to the population, and is another concept we will discuss in this chapter.

4.1. Validity

4.1.1. What is validity?

Validity asks the question, are we measuring what we want to measure? This may sound obvious but is often not that simple in educational research. Most

of the concepts we want to measure, self-concept or attitudes, for example, can't be measured directly. Self-esteem is an abstract concept, which is in some ways brought into existence by being measured. We cannot plug directly into people's heads and know what they are thinking, feeling or experiencing. It is in that sense a *latent variable*, a variable that can't be directly measured. Therefore, we need to develop instruments that measure these concepts indirectly – by using a questionnaire, for example. Every question then becomes a *manifest variable* (a variable we actually measure), designed to tease out an underlying latent concept. Creating the right measurement instrument, with the right manifest measures of the latent concept, is clearly of crucial importance, and not necessarily easy to achieve. The same is the case for concepts that may at first seem more straightforward. One of the mainstays of quantitative educational research is, of course, the achievement test. This is often used in an unproblematic way as an outcome measure in educational studies (e.g. what school characteristics affect achievement). The question, though, is what we want to measure. Often we want to make some broader comment about pupils' ability or achievement in a subject, or even more broadly, about their learning. After all, the actual goal of educational endeavour is learning, not scores on a particular achievement test. Learning, like self-esteem, cannot be measured directly. Again, to be able to do that, we would need to directly plug into people's brains and see what has actually happened there. Tests, essays and whatever other measures we use are always indirect measures of learning. Whether they are good measures is hotly disputed. Are the tests too narrow? Are they measuring higher-order skills, or only basic skills? All these questions relate directly to the validity of the test.

This means that validity is probably the single most important aspect of the design of any measurement instrument in educational research. However good our research design, or sophisticated our statistical analyses, the results will be meaningless if we aren't actually measuring what we are purporting to measure.

4.1.2. Types of validity

Validity has three distinct aspects, all of which are important: content validity, criterion validity and construct validity.

4.1.2.1. Content validity

Content validity refers to whether or not the content of the manifest variables (e.g. items of a test or questions of a questionnaire) is right to measure the latent concept (self-esteem, achievement, attitudes etc.) that we are trying to measure. For example, if we were trying to measure pupils' attitudes to school, we couldn't ask, 'How do you get on with your parents?' More difficult to determine would be an item like 'My teachers always try

to help me'. Would this be a valid measure of attitudes to school, or are attitudes to teachers something different?

Clearly, there is an important role for theory in determining content validity. The better we know our subject and how the concepts we are using are theoretically defined, the better we will be able to design an instrument that is content valid. The main judgement of whether an instrument is content valid is therefore its accordance to a theory of how the concept works and what it is.

An extensive search of the literature on the concept you want to measure will help you to achieve content validity. Asking respondents whether the instrument or test looks valid to them is also important. This is called establishing *face validity*, because respondents are judging whether the instrument looks valid to them. Setting up a panel of users and getting them to comment on your instrument while you are developing it is a good way of doing this. One problem with face validity is that lay users may not be fully cognisant of the theoretical background or subtlety of the concept, especially where you are using a psychological measure, for example. In that case, it can be useful to have a panel of experts in the field judge your instrument as well. Using a panel of experts does not mean you shouldn't also look at face validity. After all, what the people who are actually going to complete your instrument think about it is going to affect how they respond to the questions.

4.1.2.2. *Criterion validity*

As with content validity, criterion validity is closely related to theory. When you are developing a measure, you usually expect it, in theory at least, to be related to other measures, or to predict certain outcomes. For example, if we develop a new mathematics test, we would expect the scores pupils achieve on that test not to be totally unrelated to those they get on a state-mandated mathematics test.

There are two main types of criterion validity: *predictive validity* and *concurrent validity*.

Predictive validity refers to whether or not the instrument you are using predicts the outcomes you would theoretically expect it to. For example, when we select students to study on our university courses, we will use their scores on specific tests (e.g. SAT or A-level) to determine whether or not they are likely to successfully complete the course and are therefore suitable candidates. Any test we use for this purpose should therefore predict academic success. Likewise, whenever we develop a screening test for selection of employees, we expect this test to predict how well the prospective employee will do the job. Establishing whether or not this is the case will determine whether or not our measure has predictive validity.

Concurrent validity makes a less stringent assumption. The question here is whether scores on your instrument agree with scores on other factors you would expect to be related to it. For example, if you were to measure attitudes to school, you would, from theory, expect some relationship with school achievement. Likewise, when designing a measure of pupil learning in geography, you would expect there to be a relationship with scores on previously existing measures of learning in that subject.

What is needed to establish criterion validity are two things: a good knowledge of theory relating to the concept, so that we can decide what variables we can expect to be predicted by, and related to it, and a measure of the relationship between our measure and those factors. To do the latter, we need first of all to collect data on those factors from the same respondents we are measuring with our new instrument, and, secondly, to statistically measure whether there is a relationship, using techniques such as the correlation coefficient (see Chapter 8).

4.1.2.3. *Construct validity*

Construct validity is a slightly more complex issue, relating to the internal structure of an instrument and the concept it is measuring. Once again, this is related to our theoretical knowledge of the concept we want to measure. We might hypothesise that our concept or achievement measure has a number of different dimensions. For example, a test of mathematics ability might include items relating to number, shape and space, etc. We would then want to know whether all the items relate to the right dimension (e.g. an item that we have designed to measure number should measure number and not shape and space, or a combination of the two).

An example might help to clarify this. We might want to look at a measure of pupil self-concept. Desk research (looking at the literature on self-concept) suggests that this is a multidimensional construct. We can have different self-concepts in different areas. For example, I might have a positive self-concept of myself as a quantitative researcher, and a far more negative one of myself as a cook. The same goes for primary school pupils. Shavelson (1976) hypothesised that among children and adolescents seven dimensions were the most important: self-concept of school subjects, self-concept of English, self-concept of maths, self-concept of relations with peers (other children), self-concept of relations with parents, self-concept of appearance, and self-concept of athletic ability. These factors are arranged in the mind in a hierarchical manner, meaning that the three school-related factors go together to form an *academic self-concept* (e.g. 'I'm generally a good student') while the other four factors go together to form a *non-academic self-concept*. These then form the overall or global self-concept, as depicted in Figure 4.1.

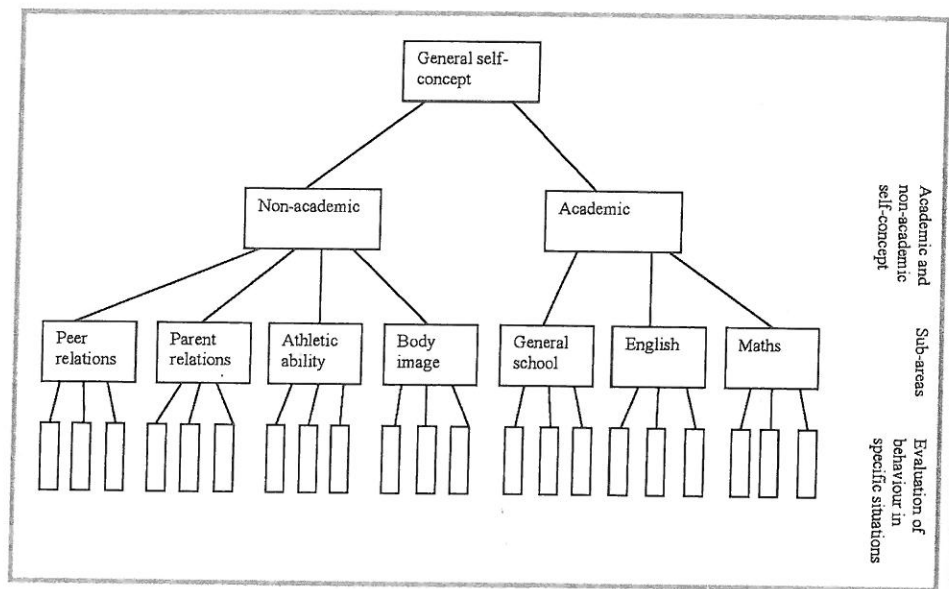


Figure 4.1 Shavelson's multifaceted, hierarchical self-concept model (Byrne & Shavelson, 1986: 476). Reproduced with permission of the authors.

If we want to measure self-concept according to this model, we will want to develop a measure that includes questions on all seven subscales. We would then want to be sure that items that were supposed to measure the peer-relations self-concept measured that subscale, and not the body-image self-concept. To do this, we can use a statistical method called confirmatory factor analysis (part of structural equation modelling), which we will look at in Chapter 11. This will tell us whether each item measures the subscale it is supposed to measure, and not any other.

Obviously, in some cases, we might not hypothesise that our construct had multiple subscales. It could be that we only wanted to measure one very specific aspect. In that case, all items would have to measure that one construct. Factor analysis would tell us whether this is the case, or whether, while we thought that all the items measure one construct, they in fact measure several. If we were to measure a certain factor by using just one item, construct validity would not be relevant.

Validity is therefore (like self-concept) a multidimensional construct. However, it is important to point out here that these different types of validity are not mutually exclusive, and that it is not sufficient to demonstrate one if the others aren't present. If we could demonstrate construct validity, but content validity was not clear, then we could not say we had a valid measure. For validity to be convincingly demonstrated, we have to see all three.

Example 4.1**Developing a self-concept scale**

For a study into the relationship between self-concept and achievement, I decided to develop a measure of self-concept. I started with a wide-ranging review of the literature. The Shavelson model of self-concept, mentioned above, was considered to be a good theoretical starting point. Seven subscales: peer relations, relations with parents, body image, physical ability, mathematics, Dutch language and general school self-concepts were hypothesised. Items were developed for each subscale, based on existing English-language scales. The scales thus developed were shown to a number of primary school teachers, and tested in a local primary school. The scale was factor analysed, and subscales were correlated with one another, achievement in maths, and a measure of peer popularity. The instrument was then changed to reflect the comments of teachers and the results of the analyses (Muijs, 1997).

The grounding in theory, and asking teachers to comment on the scale, were both meant to test the content validity of the instrument. An omission was that no panel of experts was convened. The factor analysis was designed to see whether each item measured the subscale it was supposed to measure, to look at construct validity. Finally, criterion validity was tested by seeing whether the scale was related to achievement and peer popularity. This does leave a question mark over the criterion validity of the parent relations and self-image subscales.

4.2. Reliability**4.2.1. What is reliability?**

A second element that determines the quality of our measurement instruments is *reliability*. We all know the meaning of the word *reliability* outside quantitative research. Our car is reliable if it doesn't break down. Our boy/girlfriends are unreliable because they are always late when we are supposed to meet up. In measurement, reliability is a key concept. If you were on a diet and wanted to measure your weight to see whether it was having any effect, you would want to be certain that the scales you were using were measuring your weight accurately, and would not decide to add a few pounds on one day, and take a few off the next day in a random way. The same is true if we are measuring concepts such as academic achievement or teacher behaviours in educational research.

Reliability does have a specific meaning when we are talking about statistical measurement. Basically, whenever we are measuring something, there is some element of error, called *measurement error*. Reliability then refers to the extent to which test scores are free of measurement error. Any score we get on a test or scale will have three main elements:

$$\text{score} = \text{true score} + \text{systematic error} + \text{random error}$$

The true score is what we really want to measure, the score without any error. Systematic error is error that is the same from one measurement to the next. For example, if we are measuring weight on a weighing scale, it may be that we have calibrated our scale so that it starts at 2 rather than 0 kilos. Every time you measured yourself, you would be 2 kilos too heavy. This would not be a problem if we knew what the fault was; we could simply subtract 2 kilos from every measurement we got. If we don't know what the systematic error is, our measurement will become less valid. Reliability has to do with the second part of error, unsystematic or random error. This is error that will fluctuate from one measurement to the next and that is unpredictable. This type of error is usually quite limited in scientific measurement instruments, but can be quite substantial in educational measurement. Think, for example, of a school achievement test. Many elements can cause our tests to be less reliable, resulting in random error. An item may be worded in a way that can lead to confusion, or it may be too difficult, leading to guessing. Even more random elements can intervene: the mood of the pupil when taking the test, the temperature in the room, and so on. Obviously, while both can lead to unreliability, the first set is the one we can actually do something about when developing our instrument. The second set of factors we may be able to do something about when administering our instrument, but that is an element of research design rather than instrument design.

Unreliability is clearly a problem. If we measure something unreliably, our results are untrustworthy and any conclusions tainted. Unreliable instruments will also lead relationships with other variables to be lower than if they were more reliable, thus limiting our ability to come to clear research findings. Low reliability of our instruments is one of the reasons why many of the relationships we find in educational research are low.

4.2.2. Types of reliability

Reliability, as conceptualised in quantitative research, has two main forms: *repeated measurement* and *internal consistency*.

Repeated measurement has to do with our ability to measure the same thing at different times. As mentioned above, it would not do if our instrument randomly came up with different scores every time we used it. The same instrument should come up with the same answer when used with the same respondent. In order to see whether our measures are reliable in that respect, we can simply use them with the same respondents to ascertain that the answers they give haven't changed too much. This is called the *test-retest* method. One question that follows when using the test-retest method is how much time we need to let go by before retesting. This is a

difficult question. If we leave too little time, respondents might remember how they answered last time, and simply give the same answer because of this. This is called a *carry-over* effect, and can lead to overestimating the reliability of the test. However, if we leave too long between test and retest, the respondents' attitudes or opinions might have genuinely changed, or where we are using an achievement test, the respondent may have (we would hope!) learned more in the meantime. This can lead to our underestimating the reliability of our instrument. One to two weeks is often recommended as an optimal time, though the risk of some carry-over effect remains. When we have tested and retested, we need to look at how strong the relationship is between the scores on the instrument at the two time points. To do this, we can use a correlation coefficient (see Chapter 8). This needs to be as high as possible. Above 0.7 is usually considered to offer reasonable reliability for research purposes. When we want to make a high-stakes decision on the basis of the test (such as pass or fail on an exam), we would want to have a test-retest reliability of over 0.8.

A different form of repeated measurement is *interrater reliability*. This becomes important where we use more than one judge to look at a situation, such as where we have several observers of the classroom. We would then want our observers to give the same rating to an event they had all observed (for example, if we asked three observers to observe the same lesson, and then to rate an item such as 'the teacher asks open-ended questions', we wouldn't want three different ratings!). Whether this is the case can be tested simply by doing it in practice, and then comparing the responses of all the raters.

The second form of reliability is *internal consistency reliability*. This form of reliability is only applicable to instruments that have more than one item, as it refers to how homogeneous the items of a test are, or how well they measure a single construct. When developing our self-concept scale, for example, we could first see whether the seven subscales we hypothesise exist and are measured by the variables we thought they would be (testing construct validity). Then, for each subscale, we can look at whether the items measure it in a reliable, internally homogeneous way.

There are two main ways of calculating internal consistency reliability: *split-half reliability* and *coefficient alpha*. Split-half reliability works as follows: say we have an attitude to teaching measure that consists of 10 items. First, we randomly split the test into two (for example, the even and uneven items). Then we calculate respondents' scores on each 'half test'. We can then see whether the two scores are related to one another. If they are both measuring the same thing, we would expect them to be strongly related, with a correlation coefficient of over 0.8. Coefficient alpha is another measure of internal consistency. This measure would be expected

to be over 0.7 before we can say that our test is internally consistent.

When we measure internal consistency, or test-retest reliability, we may find that our test is not reliable enough. Then we need to see whether we can pinpoint any particular item as being 'at fault'. When looking at internal consistency, we can assess how strongly each individual item is correlated with the scale score. Any items that are weakly related to the test as a whole lower our reliability, and should be removed from our instrument. When looking at test-retest reliability, we can identify items that respondents are scoring very differently on at the two times. These are causing lower reliability.

What can we do to make our instruments more reliable? A lot of this has to do with simply ensuring that the quality of questions we ask is high and unambiguous (see Chapter 3). Unambiguous and clear questions are likely to be more reliable, and the same goes for items on a rating scale for observers.

Another way to make an instrument more reliable is by measuring it with more than one item. When we use more than one item, individual errors that respondents can make when answering a single item (misreading a question, for example) cancel each other out. That is why we construct scales. In general, more items means higher reliability. We don't want to necessarily take this to extremes, though. Respondents may get bored if we keep on asking them what seem like similar questions, and are then likely to start filling out questions in an increasingly unconcentrated and haphazard way. This will increase the risk of measurement error rather than reducing it. Also, as we saw in the previous chapter, we want to keep survey instruments short, and if we use scales with loads of items, we won't be able to ask about many different things. For most attitude-type scales, somewhere between four and 10 items will lead to sufficient reliability. For achievement tests, you might want more items, due to the high-stakes nature of these tests.

A final way of making instruments more reliable is to measure a construct that is very clearly and even narrowly defined. This may in some cases conflict with validity (are we measuring our concepts too narrowly?). Obviously, we want to try to create measurements that are both reliable and valid. Remember, though, that there is not much point in creating an instrument that is reliably measuring something we don't really want to measure!

4.3. Generalisability

As we saw in Chapter 2, we often have to take a sample of our population, rather than measure the population itself. We will then do the research with our sample. When we do this, the results we find relate, strictly speaking, only to that sample. Usually, we will want to generalise our findings to

the population. When I am looking at the relationship between teacher behaviours and pupil achievement, I don't just want to be able to say something about what the 100 teachers in my sample do that affects their pupils. I would really like to say something about behaviours of teachers more generally – in other words, to generalise to the population.

4.3.1. Probability and statistical significance

Generalising to the population is not something we can automatically do, however. We saw in Chapter 2 that samples are often not totally representative of the population. Results we find in our sample might be a coincidence of that sample, rather than existing within the population. For example, we may find a relationship between the use of a reward system by teachers and pupil achievement. But if we have only observed 10 teachers, this may just be because one particularly effective teacher in our sample happens to use a reward system in her lessons.

Therefore, whenever we find a relationship in our sample, this relationship may or may not exist in our population. We would like to be able to say with a certain *probability* how likely it is that we have found a relationship in our sample when it didn't exist in the population.

If you remember, in Chapter 2, we looked at the issue of hypotheses. We said that there were two possible hypotheses, the null hypothesis and the alternative hypothesis. In this case, where we have looked at the relationship between the use of a reward system in class and the performance of pupils in that class, the hypotheses might be the following:

- The null hypothesis is that there *is no* relationship between use of a reward system and pupil performance in the population.
- The alternative hypothesis is that there *is* a relationship between use of a reward system and pupil performance in the population.

In our study, we have observed our sample of 10 teachers and found a relationship between the two. If, on the basis of this finding, we accept that there is a relationship in the population as well as in our sample, two situations can occur:

1. There is a relationship in the population. In that case, we have correctly rejected the null hypothesis.
2. There is not a relationship in the population. In that case, we have wrongly rejected the null hypothesis. This is called a *type I error*.

If, on the basis of finding a relationship in the sample, we decide that there is *not* a relationship in the population (because our relationship is only weak, for example), again two situations may occur: