CHAPTER 11

Establishing the Validity and Reliability of a Research Instrument

In this chapter you will learn about:

- The concept of validity
- Different types of validity in quantitative research
- The concept of reliability
- Factors affecting the reliability of a research instrument
- Methods of determining the reliability of an instrument in quantitative research
- Validity and reliability in qualitative research

Keywords: *concurrent validity, confirmability, construct validity, content validity, credibility, dependability, external consistency, face validity, internal consistency, reliability, transferability, validity.*

In the previous two chapters we discussed various methods of data collection in both quantitative and qualitative research. The questions asked of your respondents are the basis of your findings and conclusions. These questions constitute the ‘input’ for your conclusions (the ‘output’). This input passes through a series of steps – the selection of a sample, the collection of information, the processing of data, the application of statistical procedures and the writing of a report – and the manner in which all of these are done can affect the accuracy and quality of your conclusions. Hence, it is important for you to attempt to establish the quality of your results. As a researcher you can also be asked by others to establish the appropriateness, quality and accuracy of the procedures you adopted for finding answers to your research questions. Broadly, this concept of appropriateness and accuracy as applied to a research process is called **validity**. As inaccuracies can be introduced into a study at any stage, the concept of validity can be applied to the research process as a whole or to any of its steps: study design, sampling strategy, conclusions drawn, the statistical procedures applied or the measurement procedures used. Broadly, there are two perspectives on validity:

1. Is the research investigation providing answers to the research questions for which it was

undertaken?

2. If so, is it providing these answers using appropriate methods and procedures?

In this chapter we will discuss the concept of validity as applied to measurement procedures or the research tools used to collect the required information from your respondents.

There are prominent differences between quantitative and qualitative research in relation to the concepts of validity and reliability. Because of the defined and established structures and methods of data collection in quantitative research, the concepts of validity and reliability and the methods to determine them are well developed. However, the same is not the case in qualitative research where it would be appropriate to say that these concepts cannot be rigorously applied in the same way as they are in quantitative research because of the flexibility, freedom and spontaneity given to a researcher in the methods and procedures of data collection. It becomes difficult to establish standardisation in the method(s) of data collection in qualitative research and, hence, their validity and reliability. Despite these difficulties there are some methods which have been proposed to establish validity and reliability in qualitative research which are detailed in this chapter.

The concept of validity

To examine the concept of validity, let us take a very simple example. Suppose you have designed a study to ascertain the health needs of a community. In doing so, you have developed an interview schedule. Further suppose that most of the questions in the interview schedule relate to the attitude of the study population towards the health services being provided to them. Note that your aim was to *find out about health needs* but the interview schedule is finding out what *attitudes respondents have to the health services*; thus, the instrument is not measuring what it was designed to measure. The author has come across many similar examples among students and less skilled researchers.

In terms of measurement procedures, therefore, validity is the ability of an instrument to measure what it is designed to measure: ‘Validity is defined as the degree to which the researcher has measured what he has set out to measure’ (Smith 1991: 106). According to Kerlinger, ‘The commonest definition of validity is epitomised by the question: Are we measuring what we think we are measuring?’ (1973: 457). Babbie writes, ‘validity refers to the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration’ (1989: 133). These definitions raise two key questions:

- Who decides whether an instrument is measuring what it is supposed to measure?
- How can it be established that an instrument is measuring what it is supposed to measure?

Obviously the answer to the first question is the person who designed the study, the readership of the report and experts in the field. The second question is extremely important. On what basis do you (as a researcher), a reader as a consumer or an expert make this judgement? In the social sciences there appear to be two approaches to establishing the validity of a research instrument. These approaches are based upon either logic that underpins the construction of the research tool or statistical evidence that is gathered using information generated through the use of the instrument. Establishing validity through logic implies justification of each question in relation to the objectives of the study, whereas the statistical procedures provide hard evidence by way of calculating the coefficient of correlations between the questions and the outcome variables.

Establishing a logical link between the questions and the objectives is both simple and difficult. It is simple in the sense that you may find it easy to see a link for yourself, and difficult because your justification may lack the backing of experts and the statistical evidence to convince others. Establishing a logical link between questions and objectives is easier when the questions relate to tangible matters. For example, if you want to find out about age, income, height or weight, it is relatively easy to establish the validity of the questions, but to establish whether a set of questions is measuring, say, the effectiveness of a programme, the attitudes of a group of people towards an issue, or the extent of satisfaction of a group of consumers with the service provided by an organisation is more difficult. When a less tangible concept is involved, such as effectiveness, attitude or satisfaction, you need to ask several questions in order to cover different aspects of the concept and demonstrate that the questions asked are actually measuring it. Validity in such situations becomes more difficult to establish, and especially in qualitative research where you are mostly exploring feelings, experiences, perceptions, motivations or stories.

It is important to remember that the concept of validity is pertinent only to a particular instrument and it is an ideal state that you as a researcher aim to achieve.

Types of validity in quantitative research

There are three types of validity in quantitative research:

1. face and content validity;
2. concurrent and predictive validity;
3. construct validity.

Face and content validity

The judgement that an instrument is measuring what it is supposed to is primarily based upon the logical link between the questions and the objectives of the study. Hence, one of the main advantages of this type of validity is that it is easy to apply. Each question or item on the research instrument must have a logical link with an objective. Establishment of this link is called **face validity**. It is equally important that the items and questions cover the full range of the issue or attitude being measured. Assessment of the items of an instrument in this respect is called **content validity**. In addition, the coverage of the issue or attitude should be balanced; that is, each aspect should have similar and adequate representation in the questions or items. Content validity is also judged on the basis of the extent to which statements or questions represent the issue they are supposed to measure, as judged by you as a researcher, your readership and experts in the field. Although it is easy to present logical arguments to establish validity, there are certain problems:

- The judgement is based upon subjective logic; hence, no definite conclusions can be drawn. Different people may have different opinions about the face and content validity of an instrument.
- The extent to which questions reflect the objectives of a study may differ. If the researcher

substitutes one question for another, the magnitude of the link may be altered. Hence, the validity or its extent may vary with the questions selected for an instrument.

Concurrent and predictive validity

'In situations where a scale is developed as an indicator of some observable criterion, the scale's validity can be investigated by seeing how good an indicator it is' (Moser & Kalton 1989: 356). Suppose you develop an instrument to determine the suitability of applicants for a profession. The instrument's validity might be determined by comparing it with another assessment, for example by a psychologist, or with a future observation of how well these applicants have done in the job. If both assessments are similar, the instrument used to make the assessment at the time of selection is assumed to have higher validity. These types of comparisons establish two types of validity: **predictive validity** and **concurrent validity**. Predictive validity is judged by the degree to which an instrument can forecast an outcome. Concurrent validity is judged by how well an instrument compares with a second assessment concurrently done: 'It is usually possible to express predictive validity in terms of the correlation coefficient between the predicted status and the criterion. Such a coefficient is called a validity coefficient' (Burns 1997: 220).

Construct validity

Construct validity is a more sophisticated technique for establishing the validity of an instrument. It is based upon statistical procedures. It is determined by ascertaining the contribution of each construct to the total variance observed in a phenomenon.

Suppose you are interested in carrying out a study to find the degree of job satisfaction among the employees of an organisation. You consider status, the nature of the job and remuneration as the three most important factors indicative of job satisfaction, and construct questions to ascertain the degree to which people consider each factor important for job satisfaction. After the pre-test or data analysis you use statistical procedures to establish the contribution of each construct (status, the nature of the job and remuneration) to the total variance (job satisfaction). The contribution of these factors to the total variance is an indication of the degree of validity of the instrument. The greater the variance attributable to the constructs, the higher the validity of the instrument.

One of the main disadvantages of construct validity is that you need to know about the required statistical procedures.

The concept of reliability

We use the word 'reliable' very often in our lives. When we say that a person is reliable, what do we mean? We infer that s/he is dependable, consistent, predictable, stable and honest.

The concept of **reliability** in relation to a research instrument has a similar meaning: if a research tool is consistent and stable, hence predictable and accurate, it is said to be reliable. The greater the degree of consistency and stability in an instrument, the greater its reliability. Therefore, 'a scale or test is reliable to the extent that repeat measurements made by it under constant conditions will give the same result' (Moser & Kalton 1989: 353).

The concept of reliability can be looked at from two sides:

1. How reliable is an instrument?
2. How unreliable is it?

The first question focuses on the ability of an instrument to produce consistent measurements. When you collect the same set of information more than once using the same instrument and get the same or similar results under the same or similar conditions, an instrument is considered to be reliable. The second question focuses on the degree of inconsistency in the measurements made by an instrument – that is, the extent of difference in the measurements when you collect the same set of information more than once, using the same instrument under the same or similar conditions. Hence, the degree of inconsistency in the different measurements is an indication of the extent of its inaccuracy. This 'error' is a reflection of an instrument's unreliability. Therefore, reliability is the degree of accuracy or precision in the measurements made by a research instrument. The lower the degree of 'error' in an instrument, the higher the reliability.

Let us take an example. Suppose you develop a questionnaire to ascertain the prevalence of domestic violence in a community. You administer this questionnaire and find that domestic violence is prevalent in, say, 5 per cent of households. If you follow this with another survey using the same questionnaire on the same population under the same conditions, and discover that the prevalence of domestic violence is, say, 15 per cent, the questionnaire has not given a comparable result, which may mean it is unreliable. The less the difference between the two sets of results, the higher the reliability of the instrument.

Factors affecting the reliability of a research instrument

In the social sciences it is impossible to have a research tool which is 100 per cent accurate, not only because a research instrument cannot be so, but also because it is impossible to control the factors affecting reliability. Some of these factors are:

- **The wording of questions** – A slight ambiguity in the wording of questions or statements can affect the reliability of a research instrument as respondents may interpret the questions differently at different times, resulting in different responses.
- **The physical setting** – In the case of an instrument being used in an interview, any change in the physical setting at the time of the repeat interview may affect the responses given by a respondent, which may affect reliability.
- **The respondent's mood** – A change in a respondent's mood when responding to questions or writing answers in a questionnaire can change and may affect the reliability of that instrument.
- **The interviewer's mood** – As the mood of a respondent could change from one interview to another so could the mood, motivation and interaction of the interviewer, which could affect the responses given by respondents thereby affecting the reliability of the research instrument.
- **The nature of interaction** – In an interview situation, the interaction between the interviewer and the interviewee can affect responses significantly. During the repeat interview the responses given may be different due to a change in interaction, which could affect reliability.

- **The regression effect of an instrument** – When a research instrument is used to measure attitudes towards an issue, some respondents, after having expressed their opinion, may feel that they have been either too negative or too positive towards the issue. The second time they may express their opinion differently, thereby affecting reliability.

Methods of determining the reliability of an instrument in quantitative research

There are a number of ways of determining the reliability of an instrument and these can be classified as either external or internal consistency procedures.

External consistency procedures

External consistency procedures compare findings from two independent processes of data collection with each other as a means of verifying the reliability of the measure. The two methods of doing this are as follows:

1. **Test/retest** – This is a commonly used method for establishing the reliability of a research tool. In the test/retest (repeatability test) an instrument is administered once, and then again, under the same or similar conditions. The ratio between the test and retest scores (or any other finding, for example the prevalence of domestic violence, a disease or incidence of an illness) is an indication of the reliability of the instrument – the greater the value of the ratio, the higher the reliability of the instrument. As an equation,

   (test score)/(retest) = 1

   or

   (test score) – (retest) = 0

   A ratio of 1 shows 100 per cent reliability (no difference between test and retest) and any deviation from it indicates less reliability – the less the value of this ratio, the less the reliability of the instrument. Expressed in another way, zero difference between the test and retest scores is an indication of 100 per cent reliability. The greater the difference between scores or findings obtained from the two tests, the greater the unreliability of the instrument.

   The main advantage of the test/retest procedure is that it permits the instrument to be compared with itself, thus avoiding the sort of problems that could arise with the use of another instrument.

   The main disadvantage of this method is that a respondent may recall the responses that s/he gave in the first round, which in turn may affect the reliability of the instrument. Where an instrument is reactive in nature (when an instrument educates the respondent with respect to what the researcher is trying to find out) this method will not provide an accurate assessment of its reliability. One of the ways of overcoming this problem is to increase the time span between the two tests, but this may affect reliability for other reasons, such as the maturation of respondents and the impossibility of achieving conditions similar to those under which the questionnaire was first administered.

2. **Parallel forms of the same test** – In this procedure you construct two instruments that are intended to measure the same phenomenon. The two instruments are then administered to two similar populations. The results obtained from one test are compared with those obtained from the other. If they are similar, it is assumed that the instrument is reliable.

   The main advantage of this procedure is that it does not suffer from the problem of recall found in the test/retest procedure. Also, a time lapse between the two tests is not required. A disadvantage is that you need to construct two instruments instead of one. Moreover, it is extremely difficult to construct two instruments that are comparable in their measurement of a phenomenon. It is equally difficult to achieve comparability in the two population groups and in the two conditions under which the tests are administered.

Internal consistency procedures

The idea behind internal consistency procedures is that items or questions measuring the same phenomenon, if they are reliable indicators, should produce similar results irrespective of their number in an instrument. Even if you randomly select a few items or questions out of the total pool to test the reliability of an instrument, each segment of questions thus constructed should reflect reliability more or less to the same extent. It is based upon the logic that if each item or question is an indicator of some aspect of a phenomenon, each segment constructed will still reflect different aspects of the phenomenon even though it is based upon fewer items/questions. Hence, even if we reduce the number of items or questions, as long as they reflect some aspect of a phenomenon, a lesser number of items can provide an indication of the reliability of an instrument. The internal consistency procedure is based upon this logic. The following method is commonly used for measuring the reliability of an instrument in this way:

- **The split-half technique** – This technique is designed to correlate half of the items with the other half and is appropriate for instruments that are designed to measure attitudes towards an issue or phenomenon. The questions or statements are divided in half in such a way that any two questions or statements intended to measure the same aspect fall into different halves. The scores obtained by administering the two halves are correlated. Reliability is calculated by using the product moment correlation (a statistical procedure) between scores obtained from the two halves. Because the product moment correlation is calculated on the basis of only half the instrument, it needs to be corrected to assess reliability for the whole. This is known as *stepped-up reliability*. The stepped-up reliability for the whole instrument is calculated by a formula called the Spearman–Brown formula (a statistical procedure).

Validity and reliability in qualitative research

One of the areas of difference between quantitative and qualitative research is in the use of and the importance given to the concepts of validity and reliability. The debate centres on whether or not, given the framework of qualitative research, these concepts can or even should be applied in qualitative research. As you know, validity in the broader sense refers to the ability of a research instrument to demonstrate that it is finding out what you designed it to and reliability refers to consistency in its findings when used repeatedly. In qualitative research, as answers to research questions are explored through multiple methods and procedures which are both flexible and evolving, to ensure standardisation of research tools as well as the processes becomes difficult. As a newcomer to research you may wonder how these concepts can be applied in qualitative research when it does not use standardised and structured methods and procedures which are the bases of testing validity and reliability as defined in quantitative research. You may ask how you can ascertain the ability of an instrument to measure what it is expected to and how consistent it is when the data collection questions are neither fixed nor structured.

However, there are some attempts to define and establish validity and reliability in qualitative research. In a chapter entitled ‘Competing paradigms in qualitative research’ (pp. 105–117) in the *Handbook of Qualitative Research*, edited by Denzin and Lincoln (1994), Guba and Lincoln have suggested a framework of four criteria as a part of the constructivism paradigm paralleling ‘validity’ and ‘reliability’ in quantitative research. According to them, there are two sets of criteria ‘for judging the goodness or quality of an inquiry in constructivism paradigm’ (1994: 114). These are: *‘trustworthiness’* and *‘authenticity’*. According to Guba and Lincoln, trustworthiness in a qualitative study is determined by four indicators – **credibility**, **transferability**, **dependability** and **confirmability** – and it is these four indicators that reflect validity and reliability in qualitative research. ‘The *trustworthiness* criteria of credibility (paralleling internal validity), transferability (paralleling external validity), dependability (paralleling reliability), and confirmability (paralleling objectivity)’, according to Guba and Lincoln (1994: 114) closely relates to the concepts of validity and reliability.

Trochim and Donnelly (2007) compare the criteria proposed by Guba and Lincoln in the following table with validity and reliability as defined in quantitative research:

Traditional criteria for judging quantitative research	Alternative criteria for judging qualitative research
Internal Validity	Credibility
External Validity	Transferability
Reliability	Dependability
Objectivity	Confirmability

(Trochim and Donnelly 2007: 149)

- **Credibility** – According to Trochim and Donnelly (2007: 149), ‘credibility involves establishing that the results of qualitative research are credible or believable from the perspective of the participant in the research’. As qualitative research studies explore perceptions, experiences, feelings and beliefs of the people, it is believed that the respondents are the best judge to determine whether or not the research findings have been able to reflect their opinions and feelings accurately. Hence, credibility, which is synonymous to validity in quantitative research, is judged by the extent of respondent concordance whereby you take your findings to those who participated in your research for confirmation, congruence, validation and approval. The higher the outcome of these, the higher the validity of the study.
- **Transferability** – This ‘refers to the degree to which the results of qualitative research can be

generalized or transferred to other contexts or settings' (2007: 149). Though it is very difficult to establish transferability primarily because of the approach you adopt in qualitative research, to some extent this can be achieved if you extensively and thoroughly describe the process you adopted for others to follow and replicate.

- **Dependability** – In the framework suggested by Guba and Lincoln this is very similar to the concept of reliability in quantitative research: 'It is concerned with whether we would obtain the same results if we could observe the same thing twice' (Trochim and Donnelly 2007: 149). Again, as qualitative research advocates flexibility and freedom, it may be difficult to establish unless you keep an extensive and detailed record of the process for others to replicate to ascertain the level of dependability.
- **Confirmability** – This 'refers to the degree to which the results could be confirmed or corroborated by others' (2007: 149). Confirmability is also similar to reliability in quantitative research. It is only possible if both researchers follow the process in an identical manner for the results to be compared.

To the author's mind, to some extent, it is possible to establish the 'validity' and 'reliability' of the findings in qualitative research in the form of the model suggested by Guba and Lincoln, but its success is mostly dependent upon the identical replication of the process and methods for data collection which may not be easy to achieve in qualitative research.

Summary

One of the differences in quantitative and qualitative research is in the use of and importance attached to the concepts of validity and reliability. These concepts, their use and methods of determination are more accepted and developed in quantitative than qualitative research. The concept of validity refers to a situation where the findings of your study are in accordance with what you designed it to find out. The notion of validity can be applied to any aspect of the research process. With respect to measurement procedures, it relates to whether a research instrument is measuring what it set out to measure. In quantitative research, there are two approaches used to establish the validity of an instrument: the establishment of a logical link between the objectives of a study and the questions used in an instrument, and the use of statistical analysis to demonstrate these links. There are three types of validity in quantitative research: face and content, concurrent and predictive, and construct validity. However, the use of the concept of validity in qualitative research is debatable and controversial. In qualitative research 'credibility' as described by Guba and Lincoln seems to be the only indicator of internal validity and is judged by the degree of respondent concordance with the findings. The methods used to establish 'validity' are different in quantitative and qualitative research.

The reliability of an instrument refers to its ability to produce consistent measurements each time. When we administer an instrument under the same or similar conditions to the same or similar population and obtain similar results, we say that the instrument is 'reliable' – the more similar the results, the greater the reliability. You can look at reliability from two sides: reliability (the extent of accuracy) and unreliability (the extent of inaccuracy). Ambiguity in the wording of questions, a change in the physical setting for data collection, a respondent's mood when providing information, the nature of the interaction between interviewer and interviewee, and the regressive effect of an instrument are factors that can affect the reliability of a research instrument. In qualitative research 'reliability' is measured through 'dependability' and 'confirmability' as suggested by Guba and Lincoln.

There are external and internal consistency procedures for determining reliability in quantitative research. Test/retest and parallel forms of the same test are the two procedures that determine the external reliability of a research instrument, whereas the split-half technique is classified under internal consistency procedures. There seem to be no set procedures for determining the various indicators of validity and reliability in qualitative research.

For You to Think About

- Refamiliarise yourself with the keywords listed at the beginning of this chapter and if you are uncertain about the meaning or application of any of them revisit these in the chapter before moving on.
- Explore how the concepts of reliability and validity are applicable to research in your academic field or profession.
- Consider what strategies or procedures you could put in place to limit the affect on reliability of the following factors:
    - wording of questions;
    - physical setting;
    - respondent’s mood;
    - interviewer’s mood;
    - nature of interaction;
    - regression effect of an instrument.