

- Brugha, T. and Glover, G. (1998) 'Process and health outcomes: need for clarity in systematic reviews of case management for severe mental disorders', *Health Trends*, 30 (3), 76-9.
- Campbell, M. (1994) 'Commentary: n-of-1 trials may be useful for informed decision making', *British Medical Journal*, 309: 1044-5.
- Campbell, T. and Stanley, J. (1996) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.
- Dennett, D. (1995) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. London: Allen Lane.
- Jefferson, T., Demicheli, V. and Mugford, M. (1996) *Elementary Economic Evaluation in Health Care*. London: BMJ Publications.
- Johannessen, T. (1991) 'Controlled trials in single subjects', *British Medical Journal*, 303: 173-4.
- Kazi, M. and Wilson, J. (1996) 'Applying single-case evaluation in social work', *British Journal of Social Work*, 26: 699-717.
- March, L., Irwig, L., Schwartz, J., Simpson, J., Chock, C. and Brooks, P. (1994) 'n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis', *British Medical Journal*, 309: 1041-6.
- Newman, F. (1994) 'Disabuse of the drug metaphor: introduction', *Journal of Consulting and Clinical Psychology*, 62 (5): 941.
- Oakley, A. and Fullerton, D. (1996) 'The lamp-post of research: support or illumination?', in A. Oakley and H. Roberts (eds), *Evaluating Social Interventions*. Essex: Barnados, pp. 4-38.
- Pocock, S. (1983) *Clinical Trials: a Practical Approach*. New York: John Wiley.
- Roberts, C. and Sibbald, B. (1998) 'Randomising groups of patients', *British Medical Journal*, 316: 1898.
- Rosenthal, R. and Rubin, D. (1978) 'Interpersonal expectancy effects: the first 345 studies', *The Behavioural and Brain Sciences*, 3: 377-415.
- Sapsford, R. and Abbott, P. (1992) *Research Methods for Nurses and the Caring Professions*. Buckingham: Open University Press.
- Schultz, K., Charmers, I., Hayes, R. and Altman, D. (1995) 'Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials', *Journal of American Medical Association*, 273: 408-12.
- Schwartz, S., Flamant, R. and Lellouch, J. (1988) *Clinical Trials* (trans. M. Healy). London: Academic Press.
- Senn, S. (1997) 'Regression to the mean', *Statistical Methods in Medical Research*, 6: 99-102.
- Sheldon, B. (1983) 'The use of single case experimental designs in the evaluation of social work', *British Journal of Social Work*, 13: 477-500.
- Shepperd, S., Doll, H. and Jenkinson, C. (1997) 'Randomized controlled trials', in C. Jenkinson. (ed.), *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press. pp. 6-30.
- Shepperd, S., Harwood, D., Jenkinson, C., Gray, A., Vessey, M. and Morgan, P. (1998) 'Randomised controlled trial comparing hospital at home care with inpatient hospital care. I: three month follow up of health outcomes', *British Medical Journal*, 316: 1786-91.
- Thyer, B. (1993) 'Single-system research designs', in R. Grinnerll (ed.), *Social Work Research and Evaluation*, 4th edition. Itasca, Ill., F.E. Peacock, pp. 94-117.
- Tudor Hart, J. (1993) 'Hypertension guidelines: other diseases complicate management', *British Medical Journal*, 306: 1337.
- Watson, K. (1997) 'Economic evaluation of health care', in C. Jenkinson (ed.), *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press. pp. 129-50.
- Williams, F., Popay, J. and Oakley, A. (1999) *Welfare Research: a Critical Review*. London: UCL Press.

## CHAPTER 6

RESEARCH INSTRUMENTS IN  
EXPERIMENTAL RESEARCH

Introduction — 1 The use of standard instruments — 2 Cultural specificity and instruments — 3 Data levels and qualities — 4 Instruments and data distributions, data transformations, floor and ceiling effects — 5 Validating instruments for their reliability — 6 Reliability tests — 7 Validating the validity of instruments — 8 Questions to ask about research instruments — 9 Further reading on research instruments — References and further reading

**Introduction**


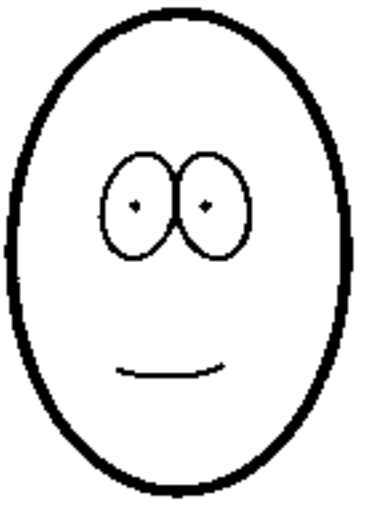
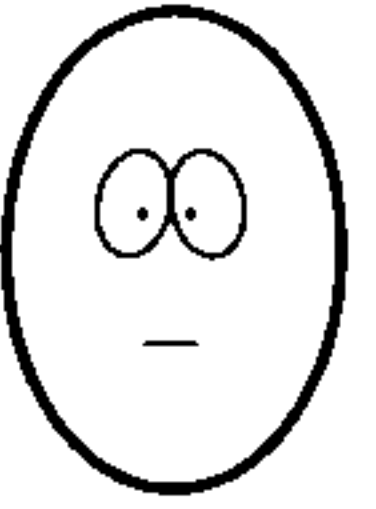


Experiments usually entail the use of some kind of data collection instrument to produce quantitative data which are then analysed statistically. The instrument used, the measurements made and the way the data are analysed shape the results. In appraising research it is important to know how the instruments and statistical tests shaped the results.

Any device that is used to aid data collection can be called an 'instrument' in research, ranging from thermometers and their associated temperature charts to questionnaires used in surveys. For research purposes, 'having a fever' may be scoring above a certain level on a thermometer, and 'being satisfied with the NHS' may be ticking a particular box on a questionnaire (see Chapter 8). Any instrument structures the data it collects. Thus looking at the design and use of such instruments is an extremely important aspect of appraising research.

Figure 6.1 gives an example of a research instrument: it shows one of the panels of the Dartmouth COOP charts used in the research by Shepperd and colleagues which is presented as the exemplar study in Chapter 3 of this volume. Table 3.1 in Chapter 3 gives data resulting from the use of the whole set of COOP charts.

Figure 6.1 gives a simple illustration of what a measurement instrument of this kind does, which is to turn ideas into numbers which can then be mathematically manipulated. It also hints at the kinds of queries this manoeuvre gives rise to. For example, what, exactly, is being measured? Nominally this is 'feelings over the past

**Figure 6.1** Dartmouth COOP chart for feelings. This is one of nine charts which make up the COOP/WONCA set. The others cover physical fitness, daily activities, social activities, pain, social support, and quality of life (Copyright Trustees of Dartmouth College COOP Project, 1989; reproduced with permission)

<b>FEELINGS</b>	
During the past 4 weeks... How much have you been bothered by emotional problems such as feeling anxious, depressed, irritable or downhearted and blue?	
Not at all	 <span style="float: right; margin-right: 10px;">1</span>
Slightly	 <span style="float: right; margin-right: 10px;">2</span>
Moderately	 <span style="float: right; margin-right: 10px;">3</span>
Quite a bit	 <span style="float: right; margin-right: 10px;">4</span>
Extremely	 <span style="float: right; margin-right: 10px;">5</span>

4 weeks'. But does the instrument capture these adequately? Will respondents understand 'feelings' as the kinds of feelings which practitioners and researchers believe are relevant to health? Questions of this sort are questions of *validity*. These also include questions about whether people respond to such instruments as intended, or perhaps as a way of making complaints, or issuing compliments to their carers.

This particular chart also presents a problem of *retrospective (or recall) bias* which arises because people reconstruct their memories according to later events, or to fit the circumstances in which they are asked about them.

There are also questions of *reliability*. Does everyone mean the same thing by 'slightly'? Would the same person confronted with the same instrument on another occasion, feeling just as good, or bad, give the same answer? Would it make a difference if the questions were asked by a key-worker, or a trained interviewer, or if the answers were given anonymously?

Questions like these arise however information is collected. The problems are just more noticeable when research instruments are used. And where instruments are used the problems are more investigable. In interview research where data are collected without the use of a research instrument (save perhaps for a checklist and a tape recorder) it is extremely difficult to know how the research process shaped the data produced, unless full transcripts of the interviews are made available (see Chapter 16). Where research instruments are used, it is possible to investigate this shaping process by testing the instruments under different circumstances. This testing is referred to as *validation*.

### 1 The use of standard instruments

Validation is very time-consuming. This encourages researchers to use instruments that have already been validated. For example, much research in health care in the UK uses instruments where people report on their own health and well-being irrespective of any particular diagnosis: *generic health measures*. Rather than inventing new instruments here most researchers choose one of the four widely used and well-validated instruments (Essink-Bot et al., 1997):

- the Nottingham Health profile – NHP (Jenkinson, 1994b; Bowling, 1995: 281–5);
- the Medical Outcomes Study 36 item Short-form Health Survey – the SF-36 (Brazier et al., 1992; Wright, 1994) (see Figure 6.2);
- the Dartmouth COOP/WONCA charts (see Figure 6.1) (Nelson et al., 1990);
- the EuroQol questionnaire (EuroQol Group, 1990; Kind et al., 1998).

In this volume, the research by Sasha Shepperd and colleagues presented in Chapter 3 uses both the COOP charts and the SF-36. There are also many instruments for recording baselines and outcomes in experiments which are specific to particular medical conditions: see, for example, the deviant behaviour rating scale used in



Chapter 2, and the World Health Organisation Angina Questionnaire used in Chapter 9.

The use of the same instruments in different pieces of research also produces results that can be compared directly with each other. Where two pieces of research on the same topic use different instruments there is always a puzzle as to whether any differences in results are real, or just the result of using different data collection instruments and measurement procedures. Chapter 4 features a systematic review of a number of different experiments on home visiting schemes and their impact on child injury rates. One of the problems encountered by the reviewers was that the different studies used different ways of measuring child injury.

Some instruments originally designed for research purposes are used to provide measures in routine health and care practice, and vice versa. For example, the *Barthel Index* used to measure the degree of assistance someone needs in order to carry out basic tasks of daily living (Bowling, 1995: 182–5) is used in both research and in routine practice. The Barthel was used by Sasha Shepperd and colleagues (Chapter 3). While two practitioners saying that their clients improved in their daily living abilities does not mean much, one practitioner saying that on average their clients improved by 10 Barthel index points, and another that theirs improved by 15 has a precise and common meaning. Thus the use of a common set of measuring instruments provides something of a common language enabling research to be applied to practice, and practice to be interpreted in the light of research findings.

## 2 Cultural specificity and instruments

Changes in linguistic habits may render any instrument out of date. There are obvious problems also of translating instruments from one language to another (including from American to British English), and of using instruments with sub-cultural groups. Instrument designers hit a particular problem here. In attempts to make instruments 'user-friendly' designers often use colloquial language (as in the COOP chart in Figure 6.1). But colloquial language is much more exclusionary than formal language for people of different generations, ethnic or dialect groups and it dates much more quickly. Consider, for example, the use of 'quite a bit', or 'blue' in Figure 6.1.

## 3 Data levels and qualities

Different instruments produce data of different kinds, or 'levels'. Box 6.1 explains what this means.

### Box 6.1 Levels of data

Different kinds of data are classified into different levels, the higher levels containing more information than the lower levels. It is possible to treat higher level data as if they were lower level data, by ignoring some of the information they contain. But it is not permissible to treat lower level data as if they were data of a higher level. Different statistical tests are appropriate for different levels of data (see Chapter 7, section 6).

- **Nominal or categorical data** – entities are classified into types and counted; for example, males and females, Yeses and Nos. No mean (average) nor median (mid-score) can be calculated: you can't have an 'average' gender. The mode or most common category is the only measure of central tendency possible.\*
- **Ordinal level data** – scores can be rank ordered, but without the distances between the ranks being measurable: for example, NHS Trust positions in a league table, clients' ranked preferences for particular kinds of services. No mean (average) can be calculated, but a median (mid-score) can be.\*
- **Interval level data and ratio level data** – scores can be placed on a scale where the difference between them can be measured precisely: for example, areas of ulcerated tissue, ages of clients, numbers of delinquent episodes. Ratio level data differ from interval level data in deriving from scales with a true zero. Both allow for the calculation of a mean (average), median (mid-point) and standard deviation.\*

Nominal and ordinal level data are often called 'qualitative data' by statisticians and medical researchers, and interval and ratio data 'quantitative data'. This is not the same quantitative/qualitative distinction which is made more generally in the methodology of the social sciences (see Chapter 16).

\* For modes, means and medians and standard deviations, see Chapter 7, section 9.

Measures of time, temperature, pressure, length, area, weight and orientation allow for the use of instruments that produce the higher level interval or ratio data. But many instruments used in health and care research do not produce higher level data, or do not do so without controversy. This is almost always so when the data concern the opinions of clients, and often when they concern the judgements of practitioners.

The COOP chart in Figure 6.1 produces data which reach the *ordinal* level. That means that scores can be put in (rank) order on a five-point scale from 'not at all' to 'extremely.' Strictly speaking, the instrument will not produce *interval* level data, since there is no way

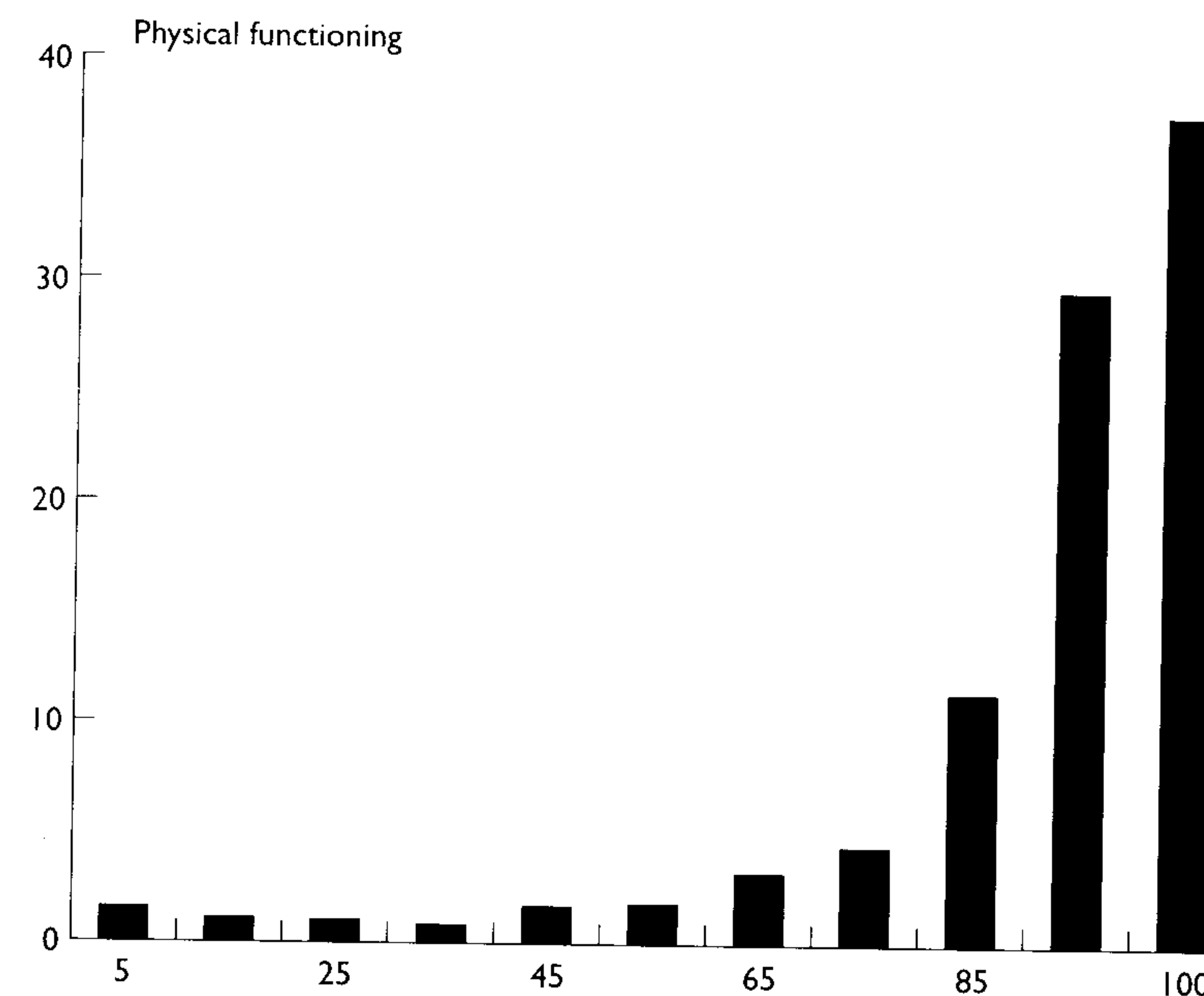
of knowing whether the gap between 'moderately' and 'slightly' is the same size as the gap between 'moderately' and 'quite a bit' – for everyone, or for anyone in particular. Taking a purist line, the appropriate measure of central tendency for ordinal level data is not the mean (average) but the median (Chapter 7, section 9). The median expresses the point below which 50 per cent of all scores and above which 50 per cent of all scores fall – the middle score. On the same principle, distributions can be described in terms of *percentiles*, where the median is the 50th percentile, where 20 per cent of scores fall below the 20th percentile and so on. Non-parametric statistics (Chapter 7, section 6) are the appropriate tests to use since these get by on comparing either the rank order of two samples, or the profiles of two samples according to the percentages of each score. Parametric tests by contrast always entail calculating a mean (average) and usually a standard deviation (see Chapter 7, section 9).

The COOP instrument (Figure 6.1), and other generic health status measures listed earlier (section 1), actually produce data no higher than ordinal. But the data are very often analysed as if they were at an interval level. Thus it will be said, perhaps, that the mean (average) score from this COOP chart for a particular population is 1.88. But that is derived from adding together '1s' which are not necessarily equal to each other nor necessarily half the value of '2s', and '4s' which are not necessarily twice the value of '2s' and so on, and dividing the total by the number of respondents. Treating data as if they were of an interval level allows for the use of the more powerful parametric statistical tests (see Chapter 7, section 6). There is controversy in general as to whether this is a sensible practice. The consensus is that sometimes it will lead to misleading results, and sometimes not, according to the instrument, the test, the sample size and whatever it is that is being measured (Pett, 1997: 32–4).

#### 4 Instruments and data distributions

Instruments also produce data that have particular shaped distributions. The ideal data distribution for statistical analysis is a *normal distribution*, meaning that when the instrument is used with a large sample, a graph of the results will take the shape of a 'bell-curve' with most results in the middle clustered around the mean/average. But many instruments used in measuring health and welfare give *skewed* distributions. This may arise when they are designed to distinguish only between degrees of unwellness, and hence tend to clump all the people who regard themselves as well together at one end of the graph. Figure 6.2 shows the distribution of scores derived from the use of the SF-36 health questionnaire with a random sample from general

Figure 6.2 A skewed distribution: distribution derived from using questions on physical functioning from the SF-36 with a random sample of general practice patients (Brazier et al., 1992: 163)



practice. The SF-36 was one of the instruments used by Shepperd and colleagues in the RCT which is the preliminary to the economic analysis in Chapter 3. From Figure 6.2 it seems that nearly 40 per cent of respondents have no problems of physical function. This kind of *skew* is common in generic health measures (section 1). There are two issues here.

#### Data transformations

One has to do with the possibilities for subjecting skewed distributions to statistical analysis, since the more powerful (parametric) statistical procedures depend on the properties of a normal (unskewed) distribution (Chapter 7, section 6). This can sometimes be solved by *transforming* the scores so that they do take something much more like a normal distribution. Thus, for example, the distribution in Figure 6.2 might take on a bell-shaped curve if instead of the raw scores, their square roots, their logarithms, or their inverses were substituted (Pett, 1997: 37, 52–4). Several of the exemplar studies use transformations for this purpose (Marshall et al., in Chapter 2; Shepperd et al., in Chapter 3; Roberts et al. in Chapter 4). (For further notes on data transformations, see Chapter 7, section 6).



The second issue has to do with the purpose of the research. Someone who was interested in mapping feelings of anxiety, depression and irritability in the general population by doing a survey would be missing the more subtle differences among 40 per cent of the population if they used the COOP chart (Figure 6.1). However, someone doing an experiment who was particularly interested in measuring changes just among people who were ill might regard the difference between a score of 2 and a score of 3 as a threshold between 'health' and 'dis-ease'. On one side of the threshold differences would be clinically interesting and important (differences among the 'suitable cases for treatment'). On the other side of the threshold any such differences might be regarded as unimportant and uninteresting.

#### *Floor and ceiling effects*

This matter is sometimes discussed in terms of *floor* and *ceiling* effects. The COOP chart produces data with a strong ceiling effect but only a small floor effect: meaning that the data *discriminate* poorly between people without serious problems, but fairly well between people with serious problems. Which is termed the 'floor' and which the 'ceiling' is an arbitrary matter.

Whether floor and ceiling effects matter depends on the uses to which an instrument is put. For example, using nominal level data to compare the effectiveness of bandaging systems for leg ulcers reduces measurement to the two categories 'healed' and 'not healed'. This produces acute floor and ceiling effects. In the study in Chapter 1, two bandaging systems were judged as equivalent, since they both produced similar rates of 'healing'. But some differences in effectiveness might be hidden below the floor, or above the ceiling of measurement if the data were nominal. In fact the authors of this study did also use additional more discriminating means of measuring and this was not the case.

### **5 Validating instruments for their reliability**

A weighing machine which kept giving different weights for the same package would be regarded as *unreliable*. It is easy enough to imagine one way of checking its reliability: keep on weighing the same package. *Test-retest* reliability means that using the instrument a second time with the same subjects will produce the same results (so long as nothing has changed between occasions of use).

In experimental work it is common for assessments by practitioners to provide baseline and/or outcome measures. There is a huge amount of research on the judgements of practitioners in most fields of health

and social care. Most of it shows that in normal practice practitioners are very *unreliable* judges in the sense that different practitioners faced with the same case make diverse judgements about the nature of the case, the diagnosis, the severity and so on. This is so even with apparently simple procedures such as taking temperatures or blood pressures (Bloor, 1978; Bloor et al., 1987; Gau and Diehl, 1982; Jenkins et al., 1985; Sackett et al., 1991). The same seems to be as true for social work as for medicine (Packman et al., 1986; Campbell, 1991). Thus, experiments that rely on practitioners making judgements 'as usual' have to be regarded with suspicion. Most experimental research using practitioner judgements attempts to enforce reliability on practitioners by providing them with a protocol or guidelines for making judgements. The protocol is, of course, the 'instrument'. It is often in a questionnaire format. An example is the MRC Needs for Care Schedule used in the study reported in Chapter 2, guiding, in this case, a psychiatrist and a psychiatric nurse in rating the care needs of subjects on entry to the experiment and after 7 and 14 months. Such instruments are validated for their reliability by testing them to see whether different practitioners using the same instrument come to the same conclusions with regard to the same cases: an *inter-rater reliability test*. Occasionally an *intra-rater reliability test* is used to see whether the same practitioner using the same instrument comes to the same conclusions when presented with the same case on two, or more, different occasions.

A different kind of reliability is *internal consistency reliability*. If an instrument has internal consistency reliability there will be a statistical correlation between those parts of the instrument allegedly measuring the same thing. In an examination that also awarded marks for grammar and spelling, for example, we would expect the same student to score much the same for this on each question answered. Many research instruments include several different ways of measuring what is allegedly the same thing, precisely for judging internal consistency.

### **6 Reliability tests**

Test-retest, inter- and intra-rater tests are usually analysed for *correlation* – sometimes called 'agreement', using a correlation coefficient called *Kappa* (Pett, 1997: 237–48), though other statistics might be used instead. Most correlation co-efficients express perfect agreement as +1 and no agreement at all as 0 and completely contrary judgements about the same matter as -1. Perfect positive or negative correlations are rare and 0.8 is regarded as a high degree of agreement, and -0.8 as a high degree of disagreement.



Correlation co-efficients are usually tested for statistical significance. This is dealt with in more detail in the next chapter (sections 1 and 2), but the issue is whether or how far an agreement might have occurred by chance. Thus for inter-rater reliability, between two judges who can only opt for 'yes' or 'no', there is already a 50 per cent chance of agreement if they merely answered at random. For three judges there is only a 25 per cent chance of agreement by chance (YYY, YYN, YNY, YNN, NYY, NYN, NNY, NNN). An 80 per cent agreement between three judges would be much more impressive than an 80 per cent agreement between two.

A Kappa, or  $\kappa$  test is often used in this context:

The weighted  $\kappa$  for the agreement between the two assessors was 0.94 for adequacy of allocation concealment, 0.51 for the extent to which the analyses were based on all randomised participants, and 0.78 for blinding. (Chapter 4, p. 40)

Here we are being told that two assessors independently assessed the quality of a set of research studies on three criteria. The extent to which they agreed is expressed by values of  $\kappa$  (Kappa). Conventionally, a  $\kappa$  of between 0.4 and 0.6 is a 'fair' level of agreement (possibly due to chance but unlikely to be so), 0.6 to 0.75 are 'good' and values greater than 0.75 are 'excellent' (Fleiss, 1971).

Internal consistency is usually measured using a co-efficient called Cronbach's alpha – the higher the alpha, the greater the consistency (Cronbach, 1951).

## 7 Validating the validity of instruments

Broadly speaking, the validity of an instrument refers to whether it measures what it is supposed to measure. There is little difficulty in agreeing that measuring changes in the area of ulcerated tissue is validly measuring whether leg ulcers are healing or not (Chapter 1). Often matters are not as simple as this. For example, the experiment featured in Chapter 2 examines which of two different ways of providing a service to people with severe mental health problems is more effective in promoting mental health. In order to measure this it is necessary for the researchers to take a position on the meaning of 'mental health', since 'mental health' and 'mental illness' are highly contested ideas. The position they adopt is a psychiatric one, and the scales they use measure matters which psychiatrists and most community mental health team members consider to be important aspects of mental health – severity of psychiatric symptoms, episodes of deviant behaviour and so on. Two of the instruments captured the

subjects' own opinions, but according to an agenda set by psychiatric ideas.

The important point here is that if the researchers had adopted a different view as to the nature of 'mental health' they would have chosen different things to measure and different instruments to measure them with, and perhaps would have produced research with different results. For example, viewed from the perspective of some self-styled 'survivors' of the mental health system, what psychiatrists would term 'deviant behaviour' could be viewed as acts of political resistance to psychiatric control. Subjects losing touch with mental health services might be viewed in terms of liberation rather than in terms of a failure of mental health care (Romme and Escher, 1993).

There are two different issues here. One is a fundamental one as to what meaning should be given to ideas such as 'health', 'illness', 'intelligence', 'social adjustment', 'equity', 'empowerment', or 'satisfaction' (as in 'consumer satisfaction'). These are all highly contested ideas, and there is no way in which research can determine their 'true' meaning. Rather researchers have to start off with some idea as to what they mean. At this fundamental level what might be a valid way of measuring mental health from a psychiatric viewpoint, would be invalid from an anti-psychiatry viewpoint, and *vice versa*.

The second issue then, is whether, once having decided what such fundamentals mean, researchers adopt a suitable way of studying them. At this level, someone antipathetic to psychiatric ideas might grudgingly concede that 'if you think that psychiatric ideas about mental health are right, then the way you are measuring it is appropriate'.

All this impacts particularly on research designed to measure 'effectiveness'. Rightly or wrongly, that term implies more than investigating the effects of doing something, and has the implication that what is 'effective' is what is also desirable. Since there may be dispute about what are more or less desirable outcomes in health and care practice, there may be disputes about what kinds of outcome measures to feature in research. One axis of the debate here is about *who* should define desirable outcomes: practitioners or service users? Much experimental research includes instruments that elicit the opinions of research subjects – as with the COOP/WONCA charts (Figure 6.1). But such information may not necessarily reflect what is of particular importance to the person from whom it is elicited. During the 1990s considerable progress was made in tempering professional judgements about desirable outcomes with the opinions of service users (see, for example, Greenhalgh et al., 1995). However, whether instruments are derived from the ideas of practitioners or from the ideas of service users, to be useful in experimental research they always have to meet standards of reliability and validity.



There are at least four different notions of validity used in the validation of research instruments. People often find it easier to understand these through the example of scholastic examinations:

- **Face validity** – the questions on the examination paper seem to be relevant to the course studied by the students and to the aims of the course they followed. The questions on the COOP chart (Figure 6.1) seem to be about the kinds of feelings which are of interest in judging morale. This is a very weak criterion for validity.
- **Content validity** – together all the questions on the examination paper cover most of the content and most of the aims of the course. Together the nine COOP charts seem to cover most dimensions of health-related quality of life, which is what they are supposed to measure.
- **Criterion validity** entails comparing results on one instrument with results on another allegedly measuring the same thing. For criterion validity there should be a correlation between what students achieved in the examination and what they achieved on continuously assessed work. The group who score most highly on the COOP charts will also be the group who are judged as most healthy by practitioners. Criterion validity is particularly important where a cheap and easy to use instrument is used instead of an expensive or intrusive investigation, as in many screening procedures.
- **Construct validity** – the results achieved from using the instrument predict those matters which the theory underlying the instrument's design says they should predict. For example, if the purpose of an examination is to differentiate students according to their ability in general terms, then those who score most highly should, as a group, be the more successful in later life. There is a self-fulfilling prophecy problem in this example, however. If the theory underlying the use of the COOP charts is that ill people with higher morale will get better quicker, then better scores on the 'feelings' chart now (Figure 6.1) should predict better scores on all the charts later. Chapters 9 and 11 deal with deprivation indices, which are validated in terms of how well they predict all the things which are associated with deprivation: death rates, morbidity rates, accident rates, low birth weights, crime rates and so on.

Judgements with regard to the last three of these criteria are usually made in terms of the strength of statistical correlations, which are explained in more detail in Chapter 10, section 10. The stronger the correlation, the better the instrument. However, it is worth noting that it is very difficult to design an instrument which is excellent in terms of all criteria of reliability and validity; a good showing on one

criterion is often achieved by a poorer showing on another. It is not enough for researchers to write that an instrument has been validated. They should say in what ways it has been validated and in terms of which criteria.

## 8 Questions to ask about research instruments

In textbook explanations the different criteria for validity above are commonly differentiated. But in real research contexts they are often difficult to distinguish from each other, and indeed, often difficult to distinguish from issues of reliability. It may be better to think of the issues here in terms of two generic questions to ask about the validity and reliability of instruments:

- What is the instrument supposed to measure?
- What evidence is there that it measures this, rather than something else?

Part 4 of this book includes a checklist of 'Questions to Ask about Data Collection Instruments'. This refers to some matters not dealt with above. For example, there are questions about whether the instrument is acceptable to the people it is used with, and whether it is appropriate for use in the context in which it is used. At first sight these may seem to be issues different from those of validity and reliability. Actually they are not, since an unacceptable instrument or one inappropriate for the context is most unlikely to produce valid results.

## 9 Further reading on research instruments

Ann Bowling's two books, *Measuring Disease* (1995) and *Measuring Health* (1991), both explain the theory of measurement and instrumentation and both provide comprehensive catalogues of a large range of research instruments used in health research, reviewing their validation history to date of publication. Crispin Jenkinson's compilation *Measuring Health and Medical Outcomes* (Jenkinson, 1994a) is particularly useful with regard to generic health status measures. His article with Hannah McGee (1997) is a lucid, though shorter, treatment of the same field. Any research paper using a validated instrument should give references to its validation pedigree, and those that do not should be regarded with some suspicion, though not with as much suspicion as those that use novel and unvalidated instruments.

## References and further reading

- Bloor, M. (1978) 'On the routinised nature of work in people-processing agencies: the case of adeno-tonsillectomy assessments in ENT outpatient clinics', in A. Davis (ed.), *Relationships between Doctors and Patients*. Farnborough: Saxon House.
- Bloor, M. (1991) 'A minor office: the variable and socially constructed character of death certification in a Scottish city', *Journal of Health and Social Behaviour*, 32: 273-87.
- Bloor, M., Samphier, M. and Prior, L. (1987) 'Artefact explanations of inequalities in health: an assessment of the evidence', *Sociology of Health and Illness*, 9(3): 231-64.
- Bowling, A. (1991) *Measuring Health: a Review of Quality of Life Measuring Scales*. Buckingham: Open University Press.
- Bowling, A. (1995) *Measuring Disease*. Buckingham: Open University Press.
- Brazier, J., Harper, R., Jones, N., O'Cathain, A., Thomas, K., Usherwood, T. and Westlake, L. (1992) 'Validating the SF-36 health survey questionnaire: a new outcome measure for primary care', *British Medical Journal*, 305: 160-4.
- Campbell, M. (1991) 'Children at risk: how different are children on child abuse registers?', *British Journal of Social Work*, 21: 259-75.
- Cronbach, I. (1951) 'Coefficient alpha and the internal consistency of tests', *Psychometrika*, 16: 297-334.
- Essink-Bot, M.-L., Krabbe, P., Bonsel, G. and Aaronson, N. (1997) 'An empirical comparison of four generic health status measures', *Medical Care*, 35 (5): 522-37.
- EuroQol Group (1990) 'EuroQol - a new facility for the measurement of health-related quality of life', *Health Policy*, 16: 199-208.
- Fleiss, J. (1971) 'Measuring nominal scale agreements among many raters', *Psychological Bulletin*, 76: 378-82.
- Gau, D. and Diehl, A. (1982) 'Disagreement among medical practitioners regarding cause of Death', *British Medical Journal*, 284: 239-40.
- Greenhalgh, J., Georgiou, A., Williams, D., Dyas, J. and Long, A. (1995) *Measuring the Outcomes of Diabetes Care*. Outcome Measurement Reviews No.4. Leeds: University of Leeds, Nuffield Institute for Health, UK Clearing House on Health Outcomes.
- Jenkins, R., Smeeton, N., Markinder, M. and Shepperd, S. (1985) 'A study of the classification of mental ill-health in general practice', *Psychological Medicine*, 15: 403-9.
- Jenkinson, C. (ed.) (1994a) *Measuring Health and Medical Outcomes*. London: UCL Press.
- Jenkinson, C., (1994b) 'Weighting for ill-health: the Nottingham Health Profile', in C. Jenkinson, (ed.), *Measuring Health and Medical Outcomes*. London: UCL Press. pp. 77-88.
- Jenkinson, C. and McGee, H. (1997) 'Patient assessed outcomes: measuring health status and quality of life', in C. Jenkinson (ed.) *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press. pp. 64-84.
- Kind, P., Dolan, P., Gudex, C. and Williams, A. (1998) 'Variations in population health status: results from a United Kingdom national questionnaire survey', *British Medical Journal*, 316: 736-40.
- Nelson, E., Langraf, J. and Hayes, R. (1990) 'The COOP Function Charts: a system to measure patient function in physicians' offices', in M. Lipkin (ed.), *Functional Status Measurement in Primary Care: Wonca Classification Committee*. New York: Springer-Verlag.
- Packman, J., Randall, J. and Jacques, N. (1986) *Who Needs Care? Social Work Decisions about Children*. Oxford: Blackwell.
- Pett, M. (1997) *Non-Parametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*. London: Sage.
- Romme, M. and Escher, S. (1993) *Accepting Voices*. London: Mind.

- Sackett, D., Haynes, R., Guyatt, G. and Tugwell, P. (1991) *Clinical Epidemiology - a Basic Science for Clinical Medicine*. London: Little, Brown and Co.
- Shepperd, S., Harwood, D., Jenkinson, C., Gray, A., Vessey, M. and Morgan, P. (1998) 'Randomised controlled trial comparing hospital at home care with inpatient hospital care: I: three month follow up of health outcomes', *British Medical Journal*, 316: 1786-91.
- Wright, L. (1994) 'The long and the short of it: the development of the SF-36 General Health Survey', in C. Jenkinson (ed.), *Measuring Health and Medical Outcomes*. London: UCL Press. pp. 89-109.