Issues in measurement

Huw TO Davies

Measurement lies at the heart of all quantitative methods in clinical, epidemiological and health services research. If the measures employed lack the essential features of validity and reliability then the conclusions drawn from empirical findings may mislead. This short article explains and explores the desirable features of measurement instruments used in health-care research.

INTRODUCTION

Measurement is central to much of health-care research. Yet if the measurement tools used are inadequate then the findings that follow will carry little conviction. This article explores the desirable features of measurement tools and explains how to assess the suitability of measures used in published studies.

TYPES OF MEASURES

In understanding measurement the first notion to grasp is that different scales can have very different basic properties. At the most basic are nominal scales: those that simply define different values that a variable can hold but say nothing about the relationship between those values. So, for example, a measure of eye colour may simply classify eyes as blue, green or brown. With nominal scales it makes no sense to say that one value is more or less than another.

More useful scales (in that they carry more information) are those that are ordinal. In ordinal scales, there exists a natural ordering so that it does make sense to say that one value is more or less than another. Examples of such scales would be standard measures of social class or deprivation, measures of educational attainment and many measures of disease severity. In all these scales there is a natural ordering of values from 'least' to 'most'.

Much of health services research employs such ordinal scales as they are easy to devise and are intuitive in interpretation. However, they are also frequently misused. Although ordinal scales by definition contain a natural ordering, there is nothing in such scales that requires the values to be evenly spaced.

For example, if a disease severity measure involved gradings of none, mild, moderate and severe, there is no reason to suppose that 'moderate' lies halfway between 'mild' and 'severe'. This limitation makes it inadmissible to calculate means and standard deviations from ordinal data: medians and inter-quartile ranges are strongly to be preferred. This issue was covered in depth in an earlier article (Davies, 1998), yet many cases can be found in the literature of the misuse of ordinal scales. For example, when data are presented as 'mean disease severity' or 'mean social class'.

Scales that have true interval properties provide more information about that which is measured, and deliver data that can be manipulated statistically. Even here there are pitfalls. Some measures have meaningful intervals but do not have a fixed and independent zero point: this makes calculating ratios problematic.

For example, at first sight it might appear that 20ºC is twice as hot as 10ºC. However, the falsity of this becomes clear when we convert to a different scale — Fahrenheit. Here 10ºC corresponds to 50ºF, and 20ºC is 68ºF. Thus it now appears that 68ºF is only 1.36 times as hot as 50ºF. This apparent paradox arises because of the essential arbitrariness of the zero in both the Fahrenheit and the Centigrade scales. When there is no fixed zero we should be careful about inferring that one value is any multiple of another. Such situations might arise for example in measures of quality of life.

MEASUREMENT PROPERTIES

Two conceptually distinct but practically intertwined issues bedevil all empirical measures: those of reliability and validity. Reliability relates to the precision of any measuring instrument (and here we are using the term 'instrument' in its broadest sense to encompass any tool used to measure, including questionnaires). A reliable measure will give the same values on repeated application so long as that which is being measured remains unchanged. Validity on the other hand reflects the extent to which the measures applied are really measuring the desired underlying properties. Clearly then, the properties of reliability and validity are matters of extent, rather than being simply present or absent.

Any given measure may be fairly valid but still unreliable. For example, measuring people's height using a simple tape measure may lack precision but still on average give a reasonable estimate of their true height. Conversely, measures may be highly reliable but nonetheless lack validity. For example, measuring height using a wrongly calibrated laser measure may give highly reproducible but nonetheless inaccurate readings.

Ideal measures have a high degree of both reliability and validity, but to some extent there are trade-offs between these: crude measures may sacrifice validity for the sake of reliability.

Dr Huw TO Davies is Reader in Health Care Policy and Management in the Department of Management, University of St Andrews, St Katharine's West, St Andrews, Fife KY16 9AL

EXPLORING RELIABILITY

Reliable measures are those that give consistent, precise and repeatable measurements. Clearly this is something that is readily tested empirically, and in that reliability is much easier to establish than validity.

One special but important aspect of measurement reliability is that of the consistency within and between raters. Suppose, for example, that clinical measures of disease severity are being estimated by clinicians as part of a research project. Then it is important to know both that each rating clinician is consistent in how he/she grades patients (intra-rater reliability) and also that there is consistency between different clinicians (inter-rater reliability). These can be tested by getting the same clinician to re-grade a batch of patients seen previously, and by comparing the ratings made by different clinicians on the same group of patients.

A high degree of agreement in either case indicates that the measurements used are reliable — however, since a certain amount of agreement is to be expected by chance alone, then this too must be taken into account in any assessment of reliability.

EXPLORING VALIDITY

Validity remains a rather more elusive and somewhat more complicated concept than reliability. Colloquially, validity is concerned with ensuring that we are measuring what we think we are measuring. Now, when some gold standard measure exists we can test validity by comparing the performance of any new measure with that of the gold standard.

For example, suppose we wish to develop a simple questionnaire to assess smoking habits, but we know that people are inaccurate reporters of their behaviour in this respect. To assess the validity of our questionnaire, we may try it out on a sample of volunteers and then compare the findings with biochemical measures of cotinine taken from saliva swabs from the same individuals (cotinine is a highly accurate measure of true exposure to tobacco smoke). If the questionnaire findings correlate well with the cotinine levels then this would reassure that the questionnaire is indeed a valid measure of smoking habits.

More often, we do not have a simple readily available gold standard against which to compare any new measuring instrument. That being so there are several different approaches to establishing the extent of validity.

Face and content validity are terms used to describe whether, on the face of it, the separate items making up a measurement tool seem to cover all the appropriate domains of interest. For example, a measure of physical function designed for assessing elderly patients which did not cover such issues as getting in and out of the bath, or climbing stairs, would be significantly lacking in content validity.

Construct validity relates to whether or not the measurement tool produces data that are consistent with known patterns. For example, a measure of angina severity should produce data that are consistent with exercise tolerance.

Closely related to this is predictive validity — this suggests that the data from a valid measure should be successful in predicting future changes. Thus, for example, a disease severity measure that did not correlate with survival would cast doubt on the validity of the instrument used to measure severity.

Finally, an important aspect of validity is sensitivity: the measures used in any given study should allow us to discriminate between groups in meaningful ways, and should be capable of detecting clinically important changes.

One crucial feature of those measures that lack validity is that their deficiencies may vary in systematic ways between groups studied. For example, measuring people's exposure to tobacco smoke is difficult and some deficiencies in validity can be expected. However, the extent of these validity problems may be expected to vary between, for example, healthy people and those recently diagnosed with lung disease. Clearly, such systematic variation may give rise to spurious findings that can mislead.

ASSESSING THE USE OF MEASURES

In assessing the appropriateness of the measures used in published reports a number of questions are germane. First: what are the scale's basic properties? That is, does the scale have ordinal, interval or ratio properties? If the scale is anything less than a true ratio scale then this limitation will need to be taken into account during analysis and interpretation.

Second: is the scale sufficiently reliable, valid and sensitive for the purpose it is being used? Answering this question is not straightforward which then leads us to ask: what efforts have the authors made to test the properties of their measures, and how successful have they been in so doing? Further, because we know that different biases may arise when measures are used with certain groups, we should ask: are there reasons to believe that the reliability/validity will differ systematically between the different study groups?

Finally we need to be aware that measures may perform well in one context (e.g. when administered by a professional in a consulting room) but may be inadequate in another (e.g. when used as part of a self-completed questionnaire). Thus we should also ask: are there contextual issues that may affect the quality of the measures obtained?

Let us now explore just one published example with measurement shortcomings. In 1996, the *Lancet* published a paper examining the use of drink and drugs by university students (Webb et al, 1996). The students' consumption was measured by way of a self-completed questionnaire distributed in lectures. There are many ways in which we can speculate that such a measure may mislead. First of all is the problem of estimating accurately an average level of a highly variable activity. It is likely that respondents' answers will be disproportionately swayed by their most recent experience. Thus the reliability of the measure is likely to be limited.

Validity too will be a problem. For example, those with large consumption of drink and drugs may under-report

for a variety of reasons, perhaps to do with social desirability. In contrast, some respondents (perhaps through bravado) may report wildly inaccurate over-estimates. The crucial point is that we have no way of knowing which of these potential biases will prevail.

Further, the context in which the instrument was used (sitting next to one's peers in a crowded lecture theatre at 9 am) may well have a marked (but again unpredictable) effect on the responses received. Unfortunately, the only reassurance given by the authors as to the measuring tool's properties was that 'discussion with students after the questionnaire sessions indicated that their reports were generally accurate'. This seems insufficient to persuade that the measures used would provide a true account of the complex behaviour of students.

CONCLUSIONS

Good measurement underlies all the quantitative methods used in health related research. No matter how good the design, if the measures are flawed then this will always undermine confidence in research findings.

Crucially, we can be misled in two ways. The use of imprecise or invalid measures may simply obscure real and important relationships leading us to miss valuable insights. For example, measures that are insensitive to change may miss small but important differences between groups. Perhaps more importantly, the use of flawed measures may purport to reveal features that are in fact only artefacts of the measures used. Measures that are invalid may produce differential responses between groups with different characteristics, leading to erroneous conclusions.

Developing good measures of complex constructs is challenging and arduous. Some helpful texts (Streiner and Norman, 1989; Oppenheim, 1992) provide considerable insights into the process, showing in greater detail how validity and reliability can be established. The complexity and uncertainty of the process is in itself a sound argument for researchers using established instruments wherever possible. At the very least, all published reports should carry a clear statement as to the soundness of the measures used. **HM**

KEY POINTS

- Good measurement underlies all quantitative research methods.
- Different measures have different scale properties: not all scales have interval or ratio properties.
- Measures should be reliable: providing consistent values when the same phenomenon is measured.
- Measures should be valid: in essence, capturing the true properties of what is being measured.
- Validity is multi-faceted and complex — both to define and to establish.
- Assessing validity may involve examining the constituent parts of the measure (content validity), identifying the consistency of known relationships (construct and predictive validity), and ensuring that small changes in reality are picked up by changes in the measures recorded (sensitivity).
- Validity and reliability may vary depending on the context within which the measures are used.
- Validity and reliability can be explored empirically — and published reports should either use previously validated scales or explain the extent to which validity and reliability have been assessed within the current study.

Davies HTO (1998) Informative presentation of summary data. *Hosp Med* **59(2):** 154–5

Oppenheim AN (1992) *Questionnaire Design, Interviewing and Attitude Measurement.* Pinter Publishers Ltd, London

Streiner DL, Norman GR (1989) *Health Measurement Scales: a Practical Guide to their Development and Use.* Oxford University Press, Oxford

Webb E, Ashton CH, Kelly P, Kamali F (1996) Alcohol and drug use in UK university students. *Lancet* **348**: 922–5