Most orthopaedic surgeons are being asked to evaluate the outcome of their practice with very little guidance as to how to accomplish this. Their approach may be simply to conduct short-term audit of perioperative complications regularly, or to implement some form of measuring instrument which can be given to all patients. They may wish to implement an outcome measure but are unable to obtain advice on which one to use. They may then apply a fashionable generic health-status measure such as the SF-36. Such types of assessment may be suitable for comparisons between disciplines, but a measure specific to a condition will normally be more appropriate for audit. Such measures are hard to find because of the wide range of conditions faced by orthopaedic surgeons. For example, in addition to pain and function, much of orthopaedics is concerned with the correction of deformity, but there have been few attempts to define outcomes for such procedures. In trauma, it is difficult to assess the state of the patient before injury. In both elective surgery and trauma the outcome may take many years to be established.

Most outcome instruments do not stand alone since a disorder must be classified in order to ensure that comparative results are for the same condition and severity. Thus, an instrument designed to measure disability due to low back pain in primary care may be more sensitive to a low level of disability than one used in a hospital setting.

What is a good outcome instrument?

Two essential requirements of an outcome instrument are that it is valid, i.e. assessing what it is supposed to, and reliable, showing a minimum of error. The measurement of these requirements is not straightforward. Most of the work in this area has been done by psychologists in studying the application of questionnaires and examinations; hence the term ‘psychometrics’ is now used to describe the study of the properties of outcome instruments. The methods of design of an instrument for measurement of outcome have been explained in detail by Streiner and Norman. In another publication, MacDowell and Newell discuss and review many such instruments and show how validity and reliability have been measured.

Content of an instrument

The items measured for outcome may be based on signs, symptoms, complications and investigations. These may be intermixed as in the Harris hip score, which uses both symptoms and signs. Outcome measures are reported either as values or categories for each dimension measured or as an overall score, termed an index.

Instruments which comprise self-reported symptoms only are much easier to use than those requiring a clinical component such as information from a radiograph, details of an operation or assessment of range of movement. Use of such data should be carefully weighed against ease of implementation and the concomitant compliance both of the user and colleagues. Incomplete data are useless. Arguments for physical assessment have been made. In the past, the belief was held that the patient’s view was of little value in measurement and that physical assessment was mandatory. The so-called ‘subjective’ tests were dismissed in favour of ‘objective’ data. In this context, ‘subjective’ means pertaining to a perceived condition within the patient’s own consciousness while ‘objective’ relates to conditions of the mind or body as viewed by another. However, it is now clear that well-designed self-reported questionnaires are very useful for determining health traits. Furthermore, many clinical signs have never been tested for reliability. The stethoscope is unreliable even in the hands of a specialist chest physician. Both inter- and intrarater reliability are very low in measuring range of movement. Self-reported questionnaires should be used whenever possible and if they are available in computerised form the information can be directly entered into a database.

Written questionnaires will usually be constructed from simple binary (yes/no) and graded questions. Graded or scaled questions are normally based on visual analogue scales (VAS), Likert scales or some form of adjectival question. Guyatt et al compared Likert with VAS scales and showed no difference in their sensitivities to change.
However, patients find that VAS questions are more difficult to understand and, in the author’s experience, they may confuse patients and delude surgeons into believing that they have continuous data.

Clearly redundant items must be excluded from any questionnaire, but they may not always be obvious since two items may ask the same question in a different way or a question may be asked about a different trait from that being measured. This heterogeneity may be identified by correlating the questions of an instrument with each other. There are various ways of doing this but Cronbach’s alpha\textsuperscript{15} is generally used. Results for homogeneity should have been published for an instrument and alpha should lie between 0.7 and 0.9.

**Validity**

Unfortunately, there is no one statistic which will tell the reader the validity of an instrument. A number of types of validity have been described by psychometricians. The three most likely to be encountered by the orthopaedic surgeon are content, criterion and construct. Content validity examines the ability of the instrument to measure all aspects of the condition for which it was designed so that it is applicable to all patients with that condition. An increase in the content will tend to decrease the reliability, but this must be viewed as secondary in importance to validity. Criterion validity establishes whether the instrument agrees with the ‘truth’. When there is no absolute truth, then construct validity should be established. This occurs, for example, in the measurement of non-specific low back pain or anterior knee pain when the patient complains of pain but no physical signs or investigations can be used to confirm either its presence or its severity. There is no direct way to test construct validity. The best approach is to ensure that the inferences made with the measure in published data are substantiated over time. The current view of psychometricians is that the most important aspect of the verification of validity is by inferences derived from the use of a measure.

**Reliability**

An instrument which is totally reliable should produce the same result each time when used under exactly the same conditions. However, in reality, results differ because of error. This may arise from the assessor (or rater), from the instrument itself, from random error or from the patient. These errors or variances may not be separated in any one particular experiment to measure reliability. Some of the issues relating to reliability in orthopaedics have been reviewed by Wright and Feinstein.\textsuperscript{16} Measures of reliability are concerned with the comparison of data. They define variances, such as those related to inter- and intrarater assessment, between methods, etc., and the words agreement, concordance, repeatability or reproducibility are often used.

The classical view is that reliability is about separating errors from the ‘true value’, but in reality the latter does not exist. A modern approach is to adopt the ‘generalisability theory’ originally developed by Cronbach in the early 1960s. This uses methods of analysis of variance to investigate the sources of error as separate entities. It develops methods for finding and reducing the main components of error. A chapter by Feldt and Brennan\textsuperscript{17} is recommended for a statistical account of reliability with many useful references to both ‘classical’ and the ‘generalisability theories’.

**Agreement**

An intraclass correlation coefficient (ICC) is most commonly used to measure agreement between or within methods or raters for continuous data. As in analyses of variance, the mean squared differences are used as sample variances to calculate an ICC. The method for calculating this will therefore depend on the design of the experiment. Results which are published showing agreement using Pearson’s product moment correlation coefficient should be treated with caution. This overestimates the agreement and is unable to distinguish data when there is a systematic error. It is always worth plotting the results to find systematic biases when comparing data. This is best achieved by placing the mean of points on the x-axis and the difference, if the comparison is between two, on the y-axis.\textsuperscript{18} This graphical technique should be viewed as complementary to the ICC.\textsuperscript{19}

If binary results are to be compared the kappa statistic is appropriate. This uses the proportion of agreement, but also adjusts for chance. For ordinal data, kappa has been extended to a weighted kappa in which the amount of disagreement is weighted. If the correct weighting is applied, then the weighted kappa statistic, and indeed kappa itself, can be directly compared with an ICC value. Lee, Koh and Ong\textsuperscript{20} suggest three measures of agreement between two methods for them to be interchangeable:

1) The lower limit of the 95% confidence interval for the ICC should be \( \geq 0.75 \).
2) There should be no systematic bias.
3) The two means should not be statistically significantly different.

**Diagnostic tests.** If the instrument is being assessed for its reliability as a clinical test, then sensitivity and specificity should be used and the precision always stated.\textsuperscript{21}

**Test-retest.** An instrument is applied and then reapplied after a short time (5 to 14 days). The results are then compared for agreement using an ICC or kappa. There is, however, often a learning bias in the retest, and such secondary application is sometimes not possible. Interitem correlations such as Cronbach’s alpha may be used as a measure of test-retest reliability.\textsuperscript{22}

**Sensitivity to change**

Sensitivity to change, or responsiveness, describes the ability of an instrument to detect changes with time. Some
measures may not need this property, for example, if death or perioperative complications are used as an outcome. However, usually, a measure is required which is sensitive to changes that occur over an interval because of the natural history of the condition or medical intervention. An instrument must be able to detect a clinically important change which may be quite small but must be clearly defined. In the experience of the author many users have no idea of what is the smallest clinically important change. Such information is vital so that the researcher can consider it when making the power calculations for the experimental design. If a small change must be detected, and the instrument is only moderately reliable, large numbers of patients are required. The ‘effect size’ is usually quoted as a measure of sensitivity to change. This index gives the distance apart of the two means in standard deviations. Cohen mentions values of 0.2, 0.5 and 0.8 as small, medium and large effect sizes, respectively.

### Synopsis

It is essential to establish an outcome measure as a valid and reliable instrument. Such assessments are designed for specific populations. A hip score may be valid for an elective total hip replacement, but not for that following trauma. It should be remembered that English words in questionnaires from North America may have different intonations to those applicable in the UK.

Once an outcome measure is chosen its scoring system should not be changed. Attempts at change or modification will make comparison with other studies valueless. Instructions which allow for missing items and incorrectly ticked boxes must be followed. The collection of outcomes and the analysis and the presentation of results takes time and therefore adequate resources and funds need to be identified.

The instrument should not be modified unless the user is prepared to go through the validation and testing of reliability again. These hurdles are high and should not be undertaken lightly. Some consequences of modifying the Oswestry Disability Questionnaire are highlighted by Fairbank and Pynsent.

An outcome measure must be easy to administer and regular feedback of aggregated results encourages compliance by staff.

### References