Measurement error and reliability testing: Application to rehabilitation

Andrea E Bialocerkowski, Peter Bragge

Aims: *Effective measurement of clinical outcomes is dependent on reliable outcome instruments. Measurement error and reliability testing are fundamental underpinnings of reliability. This article defines and illustrates sources of measurement error, outlines strategies for error minimization, and gives an overview of the types of reliability studies.*

Content: *The two main sources of measurement error: systematic bias and random error, are discussed. The three major types of reliability evaluation are then illustrated: test-retest; intra-rater and inter-rater reliability, and the relationship between reliability and validity is explained.*

Discussion and conclusions: *Quantification of measurement error is integral to determining the true effect of therapy, as quantified by outcome measurement. Interpretation of reliability data involves consideration of many factors including demographic, diagnostic and clinical characteristics of the study sample.*

Key words: ■ reliability ■ error ■ measurement studies ■ outcome measurement

Submitted 5 September; accepted for publication with minor revision following double-blind peer review 15 September 2008

Andrea E Bialocerkowski *is Senior Lecturer; and* ***Peter Bragge*** *is Lecturer, School of Physiotherapy, The University of Melbourne, 200 Berkeley Street, Carlton, VIC, 3010, AUSTRALIA*

Author for correspondence: Email: aebial@unimelb.edu.au

Outcome measurement forms an important part of therapy and rehabilitation (Huijbregts et al, 2002). Primarily, it provides information on the patient's response to treatment. However, this information can be used for other purposes, such as in goal setting, the development of management plans, justifying service provision, documenting the outcome of continuous improvement strategies, benchmarking and research (Melvin, 2001; Marshall et al, 2006). Owing to these broad potential applications, selection of appropriate outcome measures is paramount.

Guidelines for selecting outcome measures for use in clinical and research practice recommend that properties such as readability, validity, reliability, responsiveness, interpretability, time to administer and administration burden should be considered (Lohr, 2002; Terwee et al, 2007). If an outcome measure does not adequately meet these criteria, it may be unsuitable to use. Therefore, an understanding of these fundamental principles of reliability is critical in making informed decisions regarding the choice of outcome measures.

AIMS

This article presents the fundamental concepts of reliability and demonstrates how they can be applied to clinical practice. Specifically, the term 'reliability' is defined, sources of measurement error are discussed and the types of reliability are explained in a clinical context. It is outside the scope of this article to provide a detailed discussion of reliability statistics and the interpretation of reliability studies. However, readers are referred to Atkinson and Nevill (1998) and Bartlett and Frost (2008) who provide excellent summaries of this topic.

Defining reliability

Reliability is defined (Portney and Watkins, 2000) as:

> **'the extent to which a measurement is consistent and free from error'.**

'Consistency' means that the outcome measure will produce the same results on two or more occasions in a cohort whose status has not changed. 'Free from error' means that when a reliable outcome measure detects differences in patients over two more occasions, this is a result of a change in physical performance rather than measurement error.

ERRORS IN MEASUREMENT

The definition above implies that true 'reliability' is attained when a measurement is completely free from error. However, all measurements, including those derived from questionnaires and physical tests, consist of an error component (Bartlett and Frost, 2008). This means that values obtained from measurements will differ from their true value (Rothstein, 1985) and are a product of the true value plus error:

Measured value = true value + error

The true value is the value that would be gained under ideal situations when using perfect measurement techniques. Based on the above formula, subtracting the true value from the measured value would produce the error component:

Measured value – true value = error

In this equation, only the measured value is known, as it is not possible to ascertain the true value of any measurement (DeVon et al, 2007). Therefore, the error component of measurements cannot be calculated directly and requires estimation. This estimation provides information on how much of the measured value can be attributed to error and how much represents an accurate value (Portney and Watkins, 2000). This estimate is produced from reliability statistics. Knowing the magnitude of measurement error of an outcome instrument aids in interpreting whether a true change in physical performance has occurred (Atkinson and Nevill, 1998).

There are two types of measurement error: systematic bias and random error (Carmines and Zeller, 1979). Therefore:

Total measurement error = systematic bias + random error

Both of these types of errors occur in any measurement. Therefore understanding how they occur and minimizing them as much as possible is necessary to reduce total measurement error.

Systematic bias

Systematic bias refers to 'predictable errors in measurement' that occur in the same direction (under- or over-estimation) and have the same magnitude (Portney and Watkins, 2000). For example, if active wrist extension was measured on two occasions using a hand-held goniometer in ten distal radius fracture patients, systematic error would be present if the second measurement was consistently greater than the first, by a constant amount (assuming that there are no actual clinical changes in active wrist extension between the two measurement trials). One of the simplest methods to determine whether systematic bias has occurred in measurement is to graph the first and second measurements for each patient on a scatter plot. The red line in *Figure 1* illustrates theoretical perfect agreement between the two measurements (i.e. a measure of 65° on the first and second measurement, for the same patient, assuming no actual clinical change; blue diamonds on the graph). In this graph, the line of best fit through the actual measurements is located upwards and to the left of the line of perfect agreement. This means that the second measurement consistently differs from the first measurement for all patients and in this example, the second measurement is consistently 5° greater than the first.

Systematic bias may occur for a variety of reasons (Wright and Feinstein, 1992). Using the above example, fatigue after a hand strengthening programme could consistently influence measurements such that the second is always lower than the first. Alternatively, a learning effect, such as learning to extend the wrist with the fingers flexed, rather than in extension, could contribute to successive measurements being greater than the first (Atkinson and Nevill, 1998). Finally, the measurement instrument could itself contribute to systematic bias (Portney and Watkins, 2000). For example, electrogoniometers require regular calibration to reduce the likelihood of systematic measurement differences.

Because the error is predictable in nature (Portney and Watkins, 2000), systematic bias is actually more closely related to the concept of validity (an evaluation of when an instrument measures the construct that it is intended to measure) than reliability (Carmines and Zeller, 1979). Nevertheless, understanding and ascertaining the magnitude of systematic bias is an important step in evaluating an outcome measurement, as this information is invaluable in refining measurement protocols (Atkinson and Nevill, 1998). For example, a knowledge of systematic bias properties can inform recommendations regarding the number of 'training' trials required before measurement, the number of repetitions of the measurement, the period of time for recovery between measurements and the frequency of equipment calibration. Atkinson and Nevill (1998) provide an excellent

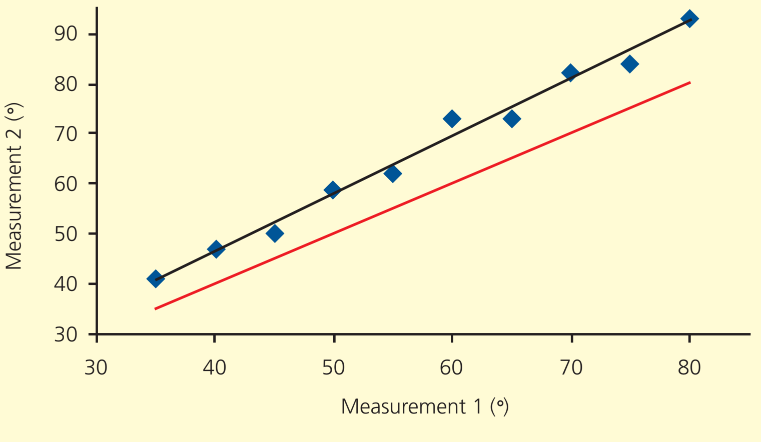

Figure 1. Systematic error: The red line is the line of perfect agreement, whereby identical values were gained for measurements one and two. The blue diamonds represent the wrist extension range of motion for ten patients. These diamonds are consistently above the red line and on average, measurement two is 5° more than measurement one.

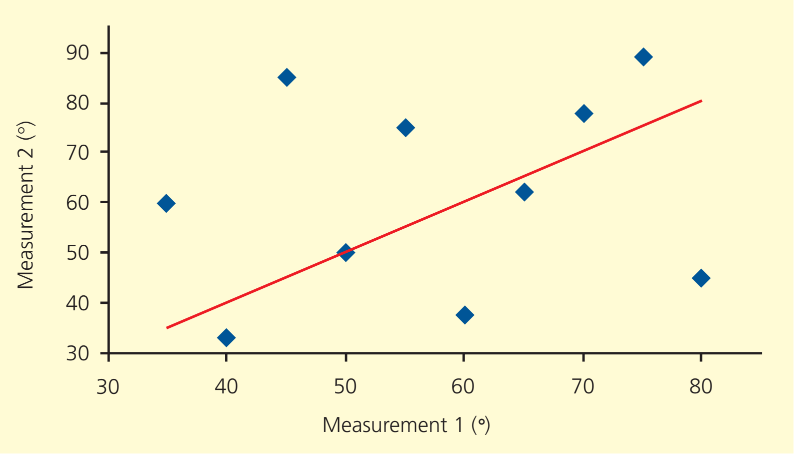

Figure 2. Random error: The red line is the line of perfect agreement, whereby identical values were gained for measurements one and two. The blue diamonds represent the wrist extension range of motion for ten patients.

summary evaluation of strategies and statistics for investigating systematic bias.

Random error

Random error, in contrast to systematic bias, is not predictable and occurs owing to chance (Carmines and Zeller, 1979). Consequently, random errors differ in their direction and magnitude between patients and occasions of testing. *Figure 2* provides a visual representation of random error. The measurements (blue diamonds) are located throughout the scatter plot, with some measurements being located above the line of perfect agreement and others below this line. In addition, the distance of each of these diamonds from the line of best fit varies, which means that the size of the error differs between patients and test occasions.

Because random errors tend to contribute more to the total error than systematic bias (Atkinson and Nevill, 1998), the evaluation of reliability focuses on determining the amount of random error in measurements (Portney and Watkins, 2000).

Sources of random error

Random error occurs for a variety of reasons, including:

- Assessor errors associated with using the equipment (such as placing a goniometer on inappropriate bony landmarks)
- Diverting from the measurement protocol (e.g. inconsistent patient position)
- Patient distraction (Wright and Feinstein, 1992).

Sources of random error for specific tests are described below.

Physical tests

In physical tests, where the clinician takes the measurement (such as the evaluation of active wrist motion), there are three sources of random error: the equipment, the patient and the clinician (Rothstein, 1985). Errors may arise from equipment that is used to take the measurement as a result of inadequate calibration or warm-up, or operation in sub-optimal conditions (Domholdt, 2005). These errors may be reduced by regular servicing and calibration of equipment, operating equipment according to the manufacturer's instructions, handling/storing equipment appropriately and using the same piece of equipment in serial measurements where possible.

The patient may also inadvertently contribute to random error, owing to lack of comprehension of what is required, lack of confidence to perform the movement or task, nervousness or distraction (Wright and Feinstein, 1992). Therefore, to minimize patient sources of random error, outcome measurement should be conducted in a quiet environment and patient instructions should be standardized, clear and simple. Allowing the patient to practice the task or movement once or twice before measurement may also reduce their nervousness.

As pain tends to impact on physical performance, this can also influence outcome measurement. Hence, the distal radius fracture patient may report increased wrist pain following mobilization, which may decrease their active range of wrist extension during post-treatment outcome measurement. Conversely, applying a hot pack may reduce wrist pain and increase active wrist extension. Consequently, it is imperative that pain intensity be evaluated in advance of outcome measurement and used to aid interpretation of measurement results.

The clinician may contribute to random error. Potentially, there may be great variability in how a clinician undertakes a measurement, even those that are simple, such as measuring active wrist extension range of motion (Wright and Feinstein, 1992). Therefore it is important to follow standardized measurement protocols where possible, such as those developed by the American Society of Hand Therapists (1992) and the American Academy of Orthopaedic Surgeons (1988) for this particular outcome measure. These protocols typically include standardized starting positions for the patient and the equipment, guidance on patient handling during the measurement, instructions to the patient, and the method of recording measurement results. To ensure consistency across patients, consensus needs to be achieved within a workplace regarding which protocols will be used for frequently performed measurements. Deviations from these protocols should be recorded in the patient's history to aid in the interpretation of measurement results. In the distal radius fracture example, the American Society of Hand Therapists (1992) goniometer protocols state that the goniometer must be placed on the volar surface of the wrist during the measurement of wrist extension.

In patients with excessive volar swelling, clinicians may choose to position the goniometer on the radial side of the wrist. This alternative method of measurement must be recorded in the patient's notes, which will assist in interpreting changes in wrist extension range of movement.

The clinician may also contribute to random error inadvertently by their level of interest in the measurement process and their level of fatigue. Moreover, their skill or competence in performing the measurement can also contribute to error (Wright and Feinstein, 1992). For example, using a hand-held goniometer is a skill that is learnt during clinical training. When learning how to use a hand-held goniometer, many students report difficulties with positioning the goniometer on standardized landmarks, providing clear and simple instructions to the patients, and reading the measurement from the goniometer. These difficulties contribute to random error and are usually overcome with practice of the technique.

Questionnaires

Self-administered questionnaires (i.e. questionnaires that are completed by patients) have two sources of error: the questionnaire and the patient (Domholdt, 2005). Items contained in the questionnaires may have ambiguous meanings or may be interpreted in different ways by different patients or by the same patient over time (Bindra et al, 2003). Clinicians can reduce random errors associated with questionnaires by providing a quiet, non-distracting environment for the patient to complete the questionnaire and removing the influence of family members or care-givers (Bindra et al, 2003). In addition, the questionnaire developers are responsible for reducing random errors associated with questionnaire items as much as possible. Questionnaire developers must define the patient characteristics for each questionnaire (such as the level of literacy required and language comprehension). During questionnaire development, psychometric evaluations are required to determine the suitability of each questionnaire item. Reliability is one such psychometric evaluation as it assesses the consistency of interpretation of the meaning of the items contained within a questionnaire.

Summary

To provide a measurement that is as close as possible to the true value, clinicians must reduce random errors as much as possible by using consistent, documented measurement procedures. When using physical tests, clinicians are responsible for reducing errors associated with the equipment they use, for patient performance and for their own performance. For self-administered questionnaires, random errors are produced by the questionnaire content and how patients interpret these items, as well as the environment in which the questionnaire is completed. Clinicians are responsible for using psychometrically sound questionnaires and administering them in a quiet environment. Moreover, clinicians should be aware of the magnitude of random error so that they can interpret their measurements appropriately (Rothstein, 1985).

TYPES OF RELIABILITY STUDIES

In addition to understanding and reducing random error, it is important to have knowledge of the different types of reliability, so appropriate reliability studies can be undertaken in the clinical setting. In general, research methodology textbooks describe three types of reliability evaluation: test-retest, intra-rater and inter-rater reliability (Rothstein, 1985).

Test-retest reliability

Test-retest reliability evaluates the consistency of the measuring instrument; that is the probability of producing the same results with repeated administration of the test. This type of reliability is particularly pertinent to situations where clinicians are not involved in the measurement process, such as self-administered questionnaires (Portney and Watkins, 2000). In this case, evaluating test-retest reliability involves a sample of patients completing the questionnaire on two occasions, under the same conditions. If the test is reliable, each patient's score will be identical across the two occasions. The period of time between occasions of testing should be carefully considered. It should be far enough apart to avoid fatigue, learning or memory effects but close enough to avoid genuine changes in the measured variable (DeVon et al, 2007). This time period needs to be conveyed by researchers in publications (DeVon et al, 2007) so that readers can evaluate whether the time between occasions of testing could influence reliability estimates. Moreover, researchers often control this aspect of the study by gaining information on the patient's perceived change between test occasions. Data from patients who report that they have not changed since the first measurement are included in reliability analyses (Portney and Watkins, 2000).

As previously mentioned, test-retest reliability of questionnaires is usually evaluated during questionnaire development. This evaluation is undertaken on patients with diagnoses that are within the target population for the questionnaire. For example, the target population of the Disabilities of the Arm, Shoulder and Hand (DASH) outcome measure is those with musculoskeletal upper limb disorders (Hudak et al, 1996). The developers evaluated its test-retest reliability in individuals with elbow problems (Solway et al, 2002) and other researchers have since evaluated reliability in individuals with a variety of upper limb disorders (Beaton et al, 2001) including those with distal radius fractures (Westphal et al, 2002). However, clinicians may wish to use this questionnaire to evaluate the function following forearm fractures other than those of the distal radius. As reliability is population specific, additional test-retest reliability evaluations are required, using a cohort of patients who have sustained such fractures. The population specificity of measurement tools means that for the same measurement tool, reliable measurements may be gained on some people but not on others (Rothstein, 1985). Thus reliability estimates provide information on the magnitude of measurement error for a specific test on a specific group of individuals (Rothstein, 1985).

Intra-rater reliability

Intra-rater reliability refers to the consistency of measurement recorded by one rater or clinician across two or more occasions (Portney and Watkins, 2000). This type of reliability is used for physical tests (such as active wrist extension range) and involves one clinician taking measurements on two occasions, under the same conditions, using the same standardized protocols and equipment (Domholdt, 2005). If the test is 100% reliable, each patient's score will be identical on both occasions of testing, assuming no clinical change in their status. In reality, errors occur even when measurements are undertaken by one clinician. Determining the magnitude of these errors is invaluable in interpreting the results of measurement, particularly whether a patient has improved, deteriorated or remained unchanged. This concept is further discussed later in the article.

The period of time between measurements is important to consider, as clinicians may be influenced by their memory of the result of their first measurement (Portney and Watkins, 2000). Moreover, the results of intra-rater reliability evaluations must be interpreted with respect to the skill-level of the clinician. Although one clinician may demonstrate acceptable reliability when evaluating active wrist extension range, this does not mean that all clinicians will be equally reliable. This is a result of differing levels of skill among clinicians in conducting the measurement, and their ability to identify and minimize potential sources of random error (Bartlett and Frost, 2008).

In the clinical and research settings, if measurements are being made by one clinician, it may be advantageous to determine the measurement error associated with these measurements. This is particularly important if there is a paucity of research evidence regarding the magnitude of measurement error produced by other clinicians with comparable skill levels, when using the same instrument on the same type of patients.

Inter-rater reliability

Variations in reliability between clinicians when using the same measurement are evaluated by assessing inter-rater reliability. Inter-rater reliability refers to the consistency of measurement recorded by two or more raters or clinicians on the same cohort of patients (Portney and Watkins, 2000). This is the probability of different clinicians producing the same results on the same patients (Domholdt, 2005). Similar to intra-rater reliability, this type of reliability is generally evaluated for physical tests (such as active wrist extension range). It involves two clinicians taking measurements on one cohort of patients, under the same conditions, using the same standardized protocols and equipment. If the test is reliable, both clinicians' scores will be identical for each individual patient.

It is important to establish the inter-rater reliability of a measure when two or more clinicians are making the same measurements on the same patients, as these measurements may significantly influence clinical decision-making. Inter-rater reliability establishes the equivalency of measurements and the magnitude of error associated when either of the clinicians are undertaking the measurement. Evidence of poor inter-rater reliability (large measurement error) may suggest deviations from the standardized test protocols and may indicate a need for further measurement training.

Interpretation of measurement results

In clinical practice, an estimation of the magnitude of the measurement error is required to interpret the results of measurement, particularly with regard to whether a patient's clinical state has altered. In therapy and rehabilitation this change is used to determine the effectiveness of an intervention (Rothstein, 1985). To state that a patient's clinical status has changed since the last measurement requires the measured change to be larger than the error associated with the measurement (Wright and Feinstein, 1992). The following example illustrates this concept. If we hypothesize that the measurement error associated with hand-held goniometry at the wrist is 5° and the patient's wrist extension was assessed as increasing by 8° we can say with some certainty that the wrist movement has increased, as the measured change exceeds that of measurement error. However, if wrist extension was assessed as increasing by 3°—a figure well below that of measurement error—there is a high probability that wrist extension range has not changed.

Relationship between reliability and validity

As stated earlier, validity is an evaluation of whether an instrument measures a construct or variable that it is intended to measure (Carmines and Zeller, 1979). It implies that a measurement is relatively free from error, i.e. it is reliable. However, if a measurement has acceptable reliability, this does not guarantee that the measurement is valid. Rather, both validity and reliability need to be evaluated during instrument development (Portney and Watkins, 2000) and considered when making measurement decisions.

The inter-related nature of validity and reliability can be illustrated by using the distal radius fracture example. A hand-held goniometer is considered to be a valid instrument for measuring wrist extension range because it evaluates angular movement. Acceptable reliability using a hand-held goniometer for the measurement of wrist extension range in distal radius fracture patients can also be demonstrated, because measurement error was minimized as much as possible by the use of a standardized measurement protocol. However, validity would be doubtful if hand sensation was assessed by evaluating wrist extension motion, as inferences about hand sensation cannot be made using this measurement. This is despite the measurement being reliable. Therefore, having an understanding of the relationship between validity and reliability is essential when considering measurement in the clinical setting.

CONCLUSIONS

Reliability is a scientific term that is frequently used in clinical practice, yet its relationship to measurement error is often not well understood. All therapy and rehabilitation measurements involve error and it is the responsibility of the clinician to reduce these errors as much as possible, particularly those that are random in nature. This is required to produce a measurement that is close as possible to its true value. Evidence regarding the test-retest, intra- and inter-rater reliability can be found for many self-administered questionnaires and physical tests. Interpretation of this evidence is dependent on an understanding of how the reliability studies were undertaken, in particular: the demographic characteristics and diagnosis of the sample, whether a sample of clinically stable patients was used, the time between measurements and the skill level of clinicians. Reliability, however, is not just a scientific phenomenon. In the clinical setting it is imperative to have an understanding of measurement errors—both those which may be attributed to the clinician and those related to the measurement instrument—so that the extent of true change in patient status can be identified (Rothstein, 1985). IJTR

Conflict of interest: none

American Academy of Orthopaedic Surgeons (1988) *Joint motion: method of measuring and recording*. Churchill Livingstone, Edinburgh

American Society of Hand Therapists (1992) Clinical assessment recommendations. The American Society of Hand Therapists, USA

Atkinson G, Nevill AM (1998) Statistical Methods For Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Med* **26**(4): 217–38

Bartlett JW, Frost C (2008) Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* **31**(4): 466–75

Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C (2001) Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. *J Hand Ther* **14**(2): 128–46

Bindra RR, Dias JJ, Heras-Palau C, Amadio PC, Chung KC, Burke FD (2003) Assessing outcome after hand surgery: the current state. *J Hand Surg [Br]* **28**(4): 289–94

Carmines E, Zeller R (1979) *Reliability and validity assessment.* Sage Publications, Beverley Hills

DeVon HA, Block ME, Moyle-Wright P et al (2007) A psychometric toolbox for testing validity and reliability. *J Nurs Scholarsh* **39**(2): 155–64

Domholdt E (2005) *Physical therapy research. Principles and applications*. WB Saunders Company, Philadelphia

Hudak P, Amadio P, Bombardier C et al (1996) Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder, and head). *Am J Ind Med* **29**(6): 602–8

Huijbregts MPJ, Myers AM, Kay TM,Gavin TS (2002) Systematic outcome measurement in clinical practice: challenges experienced by physiotherapists. *Physiother Can* **54**(1): 25

Lohr KN (2002) Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res* **11**(3): 193

Marshall S, Haywood K, Fitzpatrick R (2006) Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract* **12**(5): 559–68

Melvin J (2001) Outcomes research in rehabilitation: scope and challenges. *Am J Phys Med Rehabil* **80:** 78–82

Portney LG, Watkins MP (2000) *Foundations of clinical research. Applications to practice*. Prentice Hall Health, Upper Saddle River, New Jersey

Rothstein J (1985) *Measurement and clinical practice: theory and application*. Churchill Livingstone, New York

Solway S, Beaton D, McConnell S, Bombardier C (2002) *The DASH outcome measure user's manual.* Institute for Work and Health, Toronto

Terwee CB, Bot SDM, de Boer MR et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* **60**(1): 34–42

Westphal T, Piatek S, Schubert S, Schuschke T, Winckler S (2002) [Reliability and validity of the upper limb DASH questionnaire in patients with distal radius fractures]. *Z Orthop Ihre Grenzgeb* **140**(4): 447–51

Wright JG, Feinstein AR (1992) Improving the reliability of orthopaedic measurements. *J Bone Joint Surg Br* **74**: 287–91

KEY POINTS

- Using reliable outcome instruments is important in evaluating clinical change. Measurement error and reliability testing are key concepts underpinning outcome instrument reliability.
- There are two main types of measurement error: systematic bias and random error. There are three main types of reliability evaluation: test-retest, intra-rater and inter-rater.
- A range of strategies exist for minimizing measurement error.
- Clinicians should endeavor to find information supporting the reliability and protocols for the use of outcome measures.