

# Selecting outcome measures in sports medicine: a guide for practitioners using the example of anterior cruciate ligament rehabilitation

N P Bent,<sup>1</sup> C C Wright,<sup>1</sup> A B Rushton,<sup>1</sup> M E Batt<sup>2</sup>

<sup>1</sup> School of Health and Population Sciences, University of Birmingham, Birmingham, UK;  
<sup>2</sup> Centre for Sports Medicine, Nottingham University Hospitals, Nottingham, UK

Correspondence to:  
Mr N P Bent, School of Health and Population Sciences, University of Birmingham, 52 Pritchatts Road, Edgbaston, Birmingham B15 2TT, UK; [n.p.bent@bham.ac.uk](mailto:n.p.bent@bham.ac.uk)

Accepted 6 February 2009  
Published Online First  
17 February 2009

## ABSTRACT

Using examples from the field of anterior cruciate ligament rehabilitation, this review provides sports and health practitioners with a comprehensive, user-friendly, guide to selecting outcome measures for use with active populations. A series of questions are presented for consideration when selecting a measure: is the measure appropriate for the intended use? (appropriateness); is the measure acceptable to patients? (acceptability); is it feasible to use the measure? (feasibility); does the measure provide meaningful results? (interpretability); does the measure provide reproducible values? (reliability); does the measure assess what it is supposed to assess? (validity); can the measure detect change? (responsiveness); do substantial proportions of patients achieve the worst or best scores? (floor and ceiling effects); is the measure structured and scored correctly? (dimensionality and internal consistency); has the measure been tested with the types of patients with whom it will be used? (sample characteristics). Evaluation of the measure using these questions will assist practitioners in making their judgements.

Sports and health practitioners who are responsible for the management of injured athletes are routinely required to make decisions regarding the timing of exercise progression, the commencement of functional activities and return to competitive play.<sup>1</sup> The desire for the quickest possible return to sport must be balanced by considerations of athletes' safety and minimisation of re-injury risk. This might be accomplished by adopting an outcomes-based approach to treatment progression, in which patients must achieve specific outcomes before proceeding to more advanced levels of activity.<sup>2</sup> Such a strategy allows the quantification of an athlete's functional ability, comparison with pre-injury status and verification that an appropriate level of rehabilitation has been achieved.<sup>1</sup>

A number of outcome measures exist for monitoring progress and facilitating clinical decision-making during the rehabilitation of active individuals.<sup>3</sup> In addition, these measures might be used to determine the clinical and cost effectiveness of treatments, as well as providing benchmarks of pre-injury performance.<sup>4</sup> However, such measures are infrequently incorporated into routine practice. For example, a survey of Australian orthopaedic surgeons revealed that the majority would allow return to sport following anterior cruciate ligament (ACL) reconstruction without first assessing muscular strength.<sup>5</sup> In addition, an

investigation into professional rugby union in New Zealand found that players do not always undergo fitness testing before returning to competition.<sup>6</sup>

A possible reason for this lack of use is that selecting an outcome measure is a difficult task. Practitioners need to be familiar with the range of measures available to them, as well as published studies in which measurement properties (eg, the reliability and validity) of these measures have been assessed. Although aware of such properties, practitioners might feel insufficiently familiar with them to make definitive judgements regarding a measure's suitability. This critical review therefore aims to provide a comprehensive, yet user-friendly, guide to selecting outcome measures for use with active populations. Using illustrative examples from the field of ACL rehabilitation, it is intended to assist practitioners in judging the suitability of measures, as well as better understanding the published literature surrounding them.

## TYPES OF OUTCOME MEASURE

Rehabilitation outcome measures can be categorised, based upon their method of data acquisition, as either patient-reported or clinician-reported measures,<sup>7</sup> often referred to as subjective and objective measures, respectively. Patient-reported measures are questionnaires, such as the International Knee Documentation Committee (IKDC) subjective knee form, which contains items relating to symptoms and functional limitations experienced during activities of daily living and sports.<sup>8</sup> Clinician-reported measures incorporate performance-based tests (eg, strength and hopping ability), as well as passive, clinical tests (eg, the instrumented measurement of anterior knee laxity).<sup>9 10</sup>

Traditionally perceived as less valid than clinician-reported measures, patient-reported measures have gained favour in recent years due to their focus on issues important to patients. Nevertheless, the poor correlation between these two types of measure has prompted suggestions that they might provide different, yet complementary, information regarding a patient's functional ability, and that a combination of the two might provide the most comprehensive method of assessment.<sup>11 12</sup>

## QUESTIONS TO ASK WHEN SELECTING AN OUTCOME MEASURE

Various authors have described properties that should be considered when selecting an outcome measure.<sup>13-15</sup> Based on such recommendations, a series of questions are presented that should be

considered when selecting a patient or clinician-reported measure for use with active patients (see box 1). These are now discussed in greater detail.

### Is the measure appropriate for the intended use? (appropriateness)

An outcome measure should match the purpose for which it is used.<sup>13</sup> Known as appropriateness, this is a matter of selecting “the right tool for the job” and requires consideration of who and what is being measured, and why.

#### Who is being measured?

It is important to select a measure that has been purposely designed for the types of patients with whom it will be used.<sup>16</sup> First, a measure should be appropriate for the patient’s injury or condition. Generic measures can be used with various patient groups because they address all aspects of quality of life. Specific measures are intended for patients who share a particular feature. This could be the same injury or condition (condition-specific), injury to the same body part (site-specific), or the same signs and symptoms of injury (dimension-specific).<sup>17</sup> Examples are shown in table 1. The more condition-specific a measure, the more likely it is to measure outcomes specific to that particular condition, but the less likely it is to measure overall health and quality of life.<sup>13</sup>

A measure should also be appropriate for the patient’s activity level. For example, measures used with elite athletes will need to be capable of measuring higher levels of physical function than those used with recreational athletes. As many measures are not designed for use by highly active individuals, it is important to examine thoroughly any for appropriateness before use.

#### What is being measured?

When deciding which outcomes to measure, it is useful to consider four separate categories: body structure impairments (the injury itself); body function impairments (signs and symptoms of injury); activity limitations (inability to perform functional activities) and participation restrictions (inability to participate in life situations).<sup>21</sup> Examples are shown in table 2. The choice of which to measure should be based upon their relevance to the patient, the stage of rehabilitation and clinical judgement.

### Box 1 Questions to ask when selecting an outcome measure

- ▶ Is the measure appropriate for the intended use? (appropriateness)
- ▶ Is the measure acceptable to patients? (acceptability)
- ▶ Is it feasible to use the measure? (feasibility)
- ▶ Does the measure provide meaningful results? (interpretability)
- ▶ Does the measure provide reproducible values? (reliability)
- ▶ Does the measure assess what it is supposed to assess? (validity)
- ▶ Can the measure detect change? (responsiveness)
- ▶ Do substantial proportions of patients achieve the worst or best scores? (floor and ceiling effects)
- ▶ Is the measure structured and scored correctly? (dimensionality and internal consistency)
- ▶ Has the measure been tested with the types of patients with whom it will be used? (sample characteristics)

### Why is the measurement being taken?

The choice of measure will depend on whether it is being used for discriminative (telling patients apart) or evaluative (monitoring patient change) purposes.<sup>22</sup> Discriminative measures categorise patients based upon their scores at a particular point in time. For example, the Cincinnati knee rating system (Cincinnati system) ranks patients from “poor” to “excellent”.<sup>23</sup> Evaluative measures are used to monitor patient improvement or deterioration by comparing their scores at different time points. These measurement aims are not mutually exclusive, with many measures, including the Cincinnati system, fulfilling both.

### Is the measure acceptable to patients? (acceptability)

A measure should be acceptable to patients, ie, it should not take an unreasonable length of time to complete, or expose patients to an unacceptable level of injury risk or physical and emotional strain.<sup>13</sup> In addition, a patient-reported measure should contain clear, concise and unambiguous questions, written in easily intelligible language.<sup>24</sup> Ideally, patients’ views on these issues should be canvassed during the design of a measure; however, acceptability can be assessed by reviewing studies that have used a particular measure, for evidence of patient complaints, completion rates, and missing data. For example, in one study, 60% of ACL-deficient “non-copers” (patients with poor dynamic knee stability) refused to undertake hop testing due to fear of injury, suggesting poor acceptability for this patient group.<sup>25</sup>

### Is it feasible to use the measure? (feasibility)

The feasibility of using a measure (ie, the time and resources required to conduct, score and analyse it) is another important consideration.<sup>13</sup> For example, a questionnaire such as the Lysholm score<sup>26</sup> can be completed independently by patients, scored and analysed quickly using a calculator and incurs only photocopying costs. Conversely, the isokinetic assessment of muscular strength requires that clinicians be trained in the use of an expensive and space-consuming piece of equipment, supervise patients during familiarisation and testing sessions, and be competent in the use of computer software required for data analysis. Whether or not such considerations influence the choice of measure will depend upon the circumstances in which it is to be used and the resources available at that time.

### Does the measure provide meaningful results? (interpretability)

A measure must provide results that are meaningful to practitioners and patients. This is known as interpretability.<sup>27</sup> For discriminative purposes, a patient’s score should indicate whether they are functioning below, above, or at a normal level.<sup>28</sup> In the case of many clinician-reported measures, such as hop and strength testing, this is simply achieved through comparison with the uninvolved limb.<sup>29</sup> Although easily interpretable, such comparisons assume limb equality before injury and that the uninjured limb will be unaffected by injury and unimproved by rehabilitation. There is evidence casting doubt on these assumptions.<sup>29–31</sup>

Scores on patient-reported measures are often translated into meaningful labels, based on little more than the subjective judgements of their creators. For example, a Lysholm score higher than 94 is said to indicate a “normal” knee.<sup>32</sup> However, using the same benchmark for all patients is problematical, as what is normal for a 50-year-old office worker might be subnormal for a 20-year-old soccer player. To provide more

**Table 1** Types of outcome measure

Type	Example	Patient population
Generic	SF-36 questionnaire <sup>18</sup>	Any patient
Dimension-specific	McGill pain questionnaire <sup>19</sup>	Any patient with pain
Site-specific	IKDC form <sup>8</sup>	Any patient with knee injury
Condition-specific	ACL quality of life questionnaire <sup>20</sup>	Any patient with ACL injury

ACL, anterior cruciate ligament; IKDC, International Knee Documentation Committee subjective knee form; SF-36, 36-item Short-Form Health Survey.

meaningful comparisons, normative data should be presented so that an individual's score can be compared with uninjured people of the same age, gender and activity level.<sup>33</sup> Ideally, pre-injury performance records would be available for each individual patient. These could then act as benchmarks to be reached before return to full activity.<sup>34</sup>

For evaluative purposes, it is necessary to know whether a change in a patient's score is important or trivial.<sup>28</sup> For this reason, the minimal important difference (MID) of a measure should be considered.<sup>35</sup> This is the smallest change in score that a patient, or practitioner, perceives as important. For example, the MID of the IKDC form was found to be 11.5 points.<sup>36</sup> A patient whose score increases by less than this amount might still be improved, but not enough to be considered important.

There are several methods of estimating the MID of a measure, a discussion of which is beyond the scope of this paper but can be found elsewhere.<sup>37–39</sup> What is important for practitioners is that a justified MID is available. Unfortunately, there are many measures, such as the Cincinnati system, for which this is, as yet, not the case.

#### Does the measure provide reproducible values? (reliability)

A measure should provide similar values on repeated administrations in unchanged patients, a concept referred to as reliability.<sup>40</sup> The different types of reliability are explained in box 2.<sup>41</sup> For measures involving a practitioner (eg, clinician-reported measures), good intra and interrater reliability is important.<sup>42</sup> For measures in which no practitioner is involved (eg, patient-reported measures), good test–retest reliability is required.<sup>43</sup>

When reading the results of a reliability study, there are two questions to consider.

#### Do scores on the measure change with repeat assessments? (systematic bias)

The trend for all patients' scores to either improve or worsen between repeat assessments is termed systematic bias.<sup>44</sup> For example, in a study investigating the intrarater reliability of isokinetic testing, every patient might show improvement between their first and second assessments because of increased familiarity with the use of the dynamometer.<sup>45</sup> This is a form of systematic bias known as a learning effect.

An example of systematic bias in an interrater reliability study would be when patients undertaking hop tests all jump further when measured by a practitioner who provides greater verbal encouragement than others. An example in a test–retest reliability study would be when patients all report greater pain levels when completing a questionnaire in a cold room on their first assessment, compared with a warm room on their second assessment.

Systematic bias in a reliability study will be reflected by a difference in mean scores between assessment occasions.<sup>42</sup> If such a difference exists, and is large enough to be considered clinically

**Table 2** Outcome categories, adapted from World Health Organization<sup>21</sup>

Category	Example	Example measure(s)
Body structure impairment	ACL injury	MRI scan
Body function impairment	Muscle weakness	Isokinetic tests
Activity limitation	Reduced hop distance	Hop tests
Participation restriction	Unable to play sport	Functional fitness tests

MRI, magnetic resonance imaging.

important, then it is not worth reading further. Reliability studies should be designed to minimise systematic bias.<sup>46</sup>

#### Which reliability statistics have been calculated?

When important systematic bias is not present in a reliability study, it is appropriate to consider which reliability statistics have been reported. Two commonly utilised statistics are the intraclass correlation coefficient (ICC), and the kappa coefficient, with possible values for both ranging from 0 (no reliability) to 1 (perfect reliability).<sup>43</sup> The ICC is used for measurements on a continuous scale (eg, hop distance) and the kappa for measurements on a categorical scale (eg, mild, moderate and severe). It is recommended that a reliability of 0.7 is required when using a measure for research but a value of 0.9 is necessary when making decisions regarding individuals.<sup>47</sup> In reality, few measures are this reliable. In a recent study, only one of four hop tests had a reliability exceeding 0.9.<sup>29</sup> Using measures that fall somewhat short of this benchmark (0.7 and above) is still preferable to not using any measures at all; however, practitioners should be cautious about making important treatment decisions based on their results.<sup>48</sup>

A useful and complementary statistic to the ICC, for measurements on a continuous scale, is the standard error of measurement (SEM).<sup>43</sup> The SEM represents the amount of error associated with a measure and is expressed in the actual units of measurement. For example, a reported SEM of the single-hop test is 4.56 cm.<sup>49</sup>

The SEM can be used to estimate a range of scores that contains a patient's "true score". This is known as a confidence interval (see box 3).<sup>29</sup> The SEM can also be used to estimate the minimum detectable change (MDC)—the smallest change in an individual's score that is considered to be a true change and not measurement error (see box 4).<sup>50</sup> When estimating a confidence interval or the MDC, practitioners may choose how confident they want to be that the estimation is correct. Ninety per cent confidence is recommended when dealing with individual patients<sup>29, 50</sup> and is used for the examples shown.

An alternative approach to the MDC is Bland and Altman's limits of agreement.<sup>51</sup> For example, reported limits of agreement for the Lysholm score are  $-4.2$  to  $11.8$ ,<sup>52</sup> meaning that a patient deteriorating by less than 4.2 or improving by less than 11.8 points would be considered unchanged.

#### Does the measure assess what it is supposed to assess? (validity)

The extent to which a measure assesses what it is supposed to assess is termed validity.<sup>43</sup> There are four types of validity that practitioners need to consider.

#### Does the measure appear to be valid? (face validity)

The simple question of whether a measure appears to be valid is known as face validity.<sup>24</sup> This is an important consideration as patients are more likely to cooperate fully during assessments that they perceive to be relevant.<sup>53</sup>

**Box 2 Types of reliability**

- ▶ Intrarater reliability is the degree to which measurements taken by the same practitioner are consistent. In an intrarater reliability study, one practitioner takes measurements from the same group of patients on two or more occasions.
- ▶ Interrater reliability is the extent to which measurements taken by different practitioners are similar. In an interrater reliability study, two or more practitioners take measurements from the same group of patients on the same occasion.
- ▶ Test–retest reliability is the extent to which patients completing a measure provide consistent results. In a test–retest reliability study, the same group of patients completes a measure on two or more occasions.

**Box 3 Estimating a patient's true score: an example using the single-hop test**

- ▶ Measure the distance hopped by the patient
  - eg, 100.0 cm
- ▶ Multiply the SEM by 1.64
  - eg, SEM for the single-hop test = 4.56 cm
  - $4.56 \text{ cm} \times 1.64 = 7.5 \text{ cm}$
- ▶ Add 7.5 cm to the patient's score
  - eg,  $100.0 \text{ cm} + 7.5 \text{ cm} = 107.5 \text{ cm}$
- ▶ Subtract 7.5 cm from the patient's score
  - eg,  $100.0 \text{ cm} - 7.5 \text{ cm} = 92.5 \text{ cm}$
- ▶ The 90% confidence interval is 92.5 cm to 107.5 cm
- ▶ We can be 90% confident that the patient has hopped at least 92.5 cm but not more than 107.5 cm.

**Is the measure comprehensive? (content validity)**

The extent to which a measure covers all important aspects of the constructs (concepts such as knee symptoms or quality of life) being measured is known as content validity.<sup>4</sup> For example, a questionnaire assessing ACL injury symptoms would have poor content validity if it neglected to include questions about pain, an essential element of the construct being investigated. Content validity is only applicable to measures that comprise more than one component. For example, the Lysholm score consists of a number of separate questions,<sup>26</sup> whereas hopping ability is often measured using a battery of several different hop tests.<sup>29</sup>

When assessing a measure's content validity, practitioners should look for published evidence that its creators were thorough and systematic in deciding which components to include.<sup>54–55</sup> Authors should first specify and justify the constructs that they propose to measure and the patient groups for whom their measure is intended. Components for inclusion should then be selected on the basis of literature reviews, expert panel discussions and, because a measure should include components important to its target population, the views of patients, ascertained through interviews and focus group surveys.

**Do scores on the measure correlate with those of a "gold standard"? (criterion validity)**

A measure should correlate highly with other measures that assess the same construct and are already known to have excellent validity (gold standard measures). This is known as criterion validity, in which correlations of at least 0.7 are considered acceptable (0, no correlation; 1, perfect correlation).<sup>15</sup> For example, in a study investigating the validity of goniometry for measuring knee range of motion, correlation with the gold standard of radiographic imaging was as high as 0.99.<sup>56</sup> Because few gold standard measures exist, criterion validity is rarely assessed and practitioners will need to look for evidence of construct validity instead.<sup>53</sup>

**Does the measure relate to other measures and variables as expected? (construct validity)**

The degree to which a measure relates to other measures and variables in accordance with theoretically derived hypotheses is termed construct validity.<sup>15</sup> There are three main types of construct validity.

**Does the measure correlate well with related measures? (convergent validity)**

A measure should show correlation with other valid measures to which it is related, a concept called convergent validity.<sup>45</sup>

Because convergent validity does not involve comparison with a gold standard, very high correlations are not expected.<sup>24</sup> Instead, the extent of any anticipated correlation should be postulated and justified in advance.<sup>15</sup> For example, as hypothesised, the knee outcome survey activities of daily living scale showed a correlation greater than 0.6 with the Lysholm score.<sup>57</sup>

**Does the measure correlate poorly with unrelated measures? (divergent validity)**

As well as correlating well with related measures, a measure should not correlate too strongly with unrelated measures (ie, the correlation should be below 0.3). This is referred to as divergent validity.<sup>24</sup> For example, as hypothesised, a low correlation (0.18) was observed between Cincinnati system scores and patients' ages.<sup>23</sup>

**Can the measure detect differences between subgroups of patients? (known-groups validity)**

A measure should be able to discriminate between subgroups of patients who differ in some respect, such as age, gender, injury severity, or disability level. This is called known groups validity.<sup>58</sup> For example, as hypothesised, ACL-reconstructed patients with deteriorated articular cartilage were found to have significantly lower Cincinnati system scores than those with healthy cartilage.<sup>23</sup>

**Can the measure detect change? (responsiveness)**

Evaluative measures must be able to detect real change in a patient's condition, a property termed responsiveness.<sup>59</sup> Responsiveness studies usually involve a measure being used with patients before and after a period of treatment to see whether it can detect the changes that occur. As with construct validity, responsiveness is assessed by testing theoretically derived hypotheses.<sup>15</sup> These hypotheses are usually associated with the following three questions.

**Can the measure detect the effects of treatment?**

A measure should be able to detect the effects of treatment.<sup>60</sup> Two statistics commonly used for this purpose are the effect size (ES) and standardised response mean (SRM) (see box 5). For both statistics, values of at least 0.2, 0.5 and 0.8 indicate that small, moderate and large changes have been detected, respectively.<sup>61</sup> The magnitude of any expected treatment effect will depend on the type of treatment given and so should be postulated and justified in advance.<sup>62</sup> For example, as

## Review

**Box 4 Estimating the MDC: an example using the single-hop test**

To estimate the MDC with 90% confidence ( $MDC_{90}$ ):

- ▶ Multiply the SEM by 2.32
  - eg, SEM for the single-hop test = 4.56 cm
  - $4.56 \text{ cm} \times 2.32 = 10.6 \text{ cm}$
- ▶  $MDC_{90}$  is 10.6 cm
- ▶ If a patient's hop score improves or worsens by less than 10.6 cm, we can be 90% confident that they are unchanged.

hypothesised, the IKDC form detected a large improvement in knee-injured patients following a course of treatment, demonstrated by an ES and SRM larger than 0.8.<sup>36</sup>

**Do changes on the measure correlate with changes on related measures? (longitudinal convergent validity)**

The change recorded by a measure should show correlation with the change recorded by a related and valid measure.<sup>60</sup> This is referred to as longitudinal convergent validity.<sup>65</sup> The extent of any anticipated correlation should be postulated and justified in advance.<sup>15</sup> For example, the ability of four hop tests to detect change in patients undergoing postoperative ACL rehabilitation was assessed by correlating change in hop scores with change on a patient-reported measure (the lower extremity functional scale).<sup>29</sup> A correlation of 0.5 was prespecified as evidence of good responsiveness but was not achieved.

**Can the measure discriminate between subgroups of patients who change by different amounts? (longitudinal known-groups validity)**

A measure should be able to discriminate between identifiable subgroups of patients who change by different amounts.<sup>60</sup> This is known as longitudinal known-groups validity.<sup>65</sup> For example, as hypothesised, the IKDC form detected greater improvement in patients undergoing treatment for ACL injury than for osteoarthritis.<sup>36</sup>

**Do substantial proportions of patients achieve the worst or best scores? (floor and ceiling effects)**

On some measures, particularly patient-reported measures, it is possible to achieve a worst possible or best possible score. If substantial proportions of patients (15–20%) achieve the worst possible score, floor effects are said to be present, indicating that the measure is too difficult. If substantial proportions of patients achieve the best possible score, ceiling effects are present and the measure is too easy.<sup>15, 33</sup> For example, 37% of patients awaiting ACL reconstruction achieved the worst possible score on the Cincinnati system's sports function

**Box 5 Effect size and standardised response mean**

$$\text{Effect size} = \frac{\text{Mean change in scores}}{\text{Standard deviation of baseline scores}}$$

$$\text{Standardised response mean} = \frac{\text{Mean change in scores}}{\text{Standard deviation of change in scores}}$$

subscales, and 39% achieved the best possible score at final postoperative follow-up.<sup>23</sup>

When using a measure with sporting individuals, ceiling effects are of particular concern.<sup>3</sup> As their normal level of physical function is very high, these patients might achieve the best possible score on a measure and be deemed "normal" well before reaching full fitness. In addition, no further improvements in function would be detectable.

**Is the measure structured and scored correctly? (dimensionality and internal consistency)**

Measures comprising more than one component should be structured and scored in a way that is consistent with the number of constructs they measure (their dimensionality).<sup>14</sup> Measures that assess only one construct, such as the IKDC form,<sup>8</sup> are called unidimensional and scores from their components can be summed together to form an overall score.<sup>13</sup> Tests that measure more than one construct are called multidimensional; their components are grouped into distinct, unidimensional sections, each measuring a single construct and with its own total score.<sup>14</sup> For example, the knee injury and osteoarthritis outcome score comprises five distinct subscales: "pain", "symptoms", "activities of daily living", "sport and recreation" and "quality of life".<sup>64</sup> The dimensionality of a measure can be assessed through statistical techniques such as factor analysis or Rasch analysis.<sup>14</sup> For example, factor analysis demonstrated that all questions on the IKDC form were measuring a single construct.<sup>8</sup>

When a measure's dimensionality has been determined, there is one further consideration. All components within a section should be measuring the same construct and therefore be highly correlated with one another. This concept is called internal consistency and is usually measured using a statistic called Cronbach's alpha.<sup>65</sup> A value of 0.7 is recommended for research purposes but a value of 0.9 for use with individuals.<sup>47</sup> For example, Cronbach's alpha for the IKDC form was 0.92.<sup>8</sup>

**Has the measure been tested with the types of patients with whom it will be used? (sample characteristics)**

When reading the results of a study in which the measurement properties of a measure have been investigated, it is essential to note the characteristics of the patients involved. First, it is important that the sample size is adequate. A measure may appear reliable, but this conclusion might be dubious if only a small number of patients have been tested. Therefore, a study's sample size should be justified, preferably using a power calculation.<sup>66</sup>

Second, it is important to consider the types of patients employed in a study, as all measurement properties discussed in this article are population-specific.<sup>45</sup> For example, a measure that is valid for ACL-injured patients might not be valid for posterior cruciate ligament-injured patients. In addition, a measure that is reliable for recreational athletes might not be reliable for elite athletes. Therefore, when possible, a measure should be selected that has sound measurement properties for the types of patients with whom it will be used.

**CONCLUSION**

A series of questions have been presented for sports and health practitioners to consider when selecting outcome measures for use with active patients. Practitioners must judge whether a measure is appropriate, acceptable, feasible, interpretable, reliable, valid, responsive, free of floor and ceiling effects and

### What is already known on this topic

Numerous outcome measures exist for use with active patients but are infrequently used in routine clinical practice. Although aware of measurement properties such as reliability and validity, sports and health practitioners might feel insufficiently familiar with them to make definitive judgements regarding an outcome measure's suitability for their patients.

### What this study adds

This paper provides sports and health practitioners with a series of questions that should be asked when selecting an outcome measure. By considering these questions, practitioners will be better able to judge the measures available to them and select those most suitable for their patients.

structured correctly, for use with the types of patients of interest to them. Some of these judgements, such as whether it is feasible to use a measure, can be made based on familiarity with the measure itself. Others, such as whether a measure is reliable, must be based on published evidence. This article is intended to assist practitioners with these judgements.

Although knowledgeable in how to evaluate an outcome measure, selecting one might still seem like a daunting task. There are a large number of measures available, with one review identifying 16 different questionnaires intended for use with knee-injured patients alone.<sup>67</sup> Lack of confidence when selecting outcome measures might be a barrier that prevents their clinical use despite recognised benefits for patient management. However, practitioners should remember that selecting an outcome measure is a skill that needs to be practised like any other and will improve with use. Utilisation of outcome measures should increase as practitioners become more comfortable with evaluating them.

**Competing interests:** None.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

### REFERENCES

- Fuller CW, Walker J. Quantifying the functional rehabilitation of injured football players. *Br J Sports Med* 2006;**40**:151–7.
- Myer GD, Paterno MV, Ford KR, et al. Rehabilitation after anterior cruciate ligament reconstruction: criteria-based progression through the return-to-sport phase. *J Orthop Sports Phys Ther* 2006;**36**:385–402.
- Denegar CR, Vela LI, Evans TA. Evidence-based sports medicine: outcomes instruments for active populations. *Clin Sports Med* 2008;**27**:339–51.
- Ware JE Jr, Brook RH, Davies AR, et al. Choosing measures of health-status for individuals in general populations. *Am J Public Health* 1981;**71**:620–25.
- Feller JA, Cooper R, Webster KE. Current Australian trends in rehabilitation following anterior cruciate ligament reconstruction. *Knee* 2002;**9**:121–6.
- Beardmore AL, Handcock PJ, Rehrer NJ. Return-to-play after injury: practices in New Zealand rugby union. *Phys Ther Sport* 2005;**6**:24–30.
- Irrgang JJ, Lubowitz JH. Measuring arthroscopic outcome. *Arthroscopy* 2008;**24**:718–22.
- Irrgang JJ, Anderson AF, Boland AL, et al. Development and validation of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2001;**29**:600–13.
- Mattacola CG, Perrin DH, Gansnedder BM, et al. Strength, functional outcome, and postural stability after anterior cruciate ligament reconstruction. *J Athl Train* 2002;**37**:262–8.
- Daniel DM, Malcom LL, Losse G, et al. Instrumented measurement of anterior laxity of the knee. *J Bone Joint Surg Am* 1985;**67A**:720–6.
- Neeb TB, Aufdenkampe G, Wagener JH, et al. Assessing anterior cruciate ligament injuries: the association and differential value of questionnaires, clinical tests, and functional tests. *J Orthop Sports Phys Ther* 1997;**26**:324–31.

- Pantano KJ, Irrgang JJ, Burdett R, et al. A pilot study on the relationship between physical impairment and activity restriction in persons with anterior cruciate ligament reconstruction at long-term follow-up. *Knee Surg Sports Traumatol Arthrosc* 2001;**9**:369–78.
- Fitzpatrick R, Davey C, Buxton MJ, et al. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**:i–iv, 1–74.
- Pesudovs K, Burr JM, Harley C, et al. The development, assessment, and selection of questionnaires. *Optom Vis Sci* 2007;**84**:663–74.
- Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.
- Jette AM. Using health-related quality of life measures in physical therapy outcomes research. *Phys Ther* 1993;**73**:528–37.
- Guyatt GH, Veldhuyzen Van Zanten SJ, Feeny DH, et al. Measuring quality of life in clinical trials: a taxonomy and review. *Can Med Assoc J* 1989;**140**:1441–8.
- Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;**30**:473–83.
- Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. *Pain* 1975;**1**:277–99.
- Mohtadi N. Development and validation of the quality of life outcome measure (questionnaire) for chronic anterior cruciate ligament deficiency. *Am J Sports Med* 1998;**26**:350–9.
- World Health Organization. *International classification of functioning, disability and health: ICF*. Geneva: WHO, 2001.
- Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis* 1987;**40**:171–8.
- Barber-Westin SD, Noyes FR, McCloskey JW. Rigorous statistical reliability, validity, and responsiveness testing of the Cincinnati Knee Rating System in 350 subjects with uninjured, injured, or anterior cruciate ligament-reconstructed knees. *Am J Sports Med* 1999;**27**:402–16.
- Streiner DL. A checklist for evaluating the usefulness of rating-scales. *Can J Psychiatry* 1993;**38**:140–8.
- Rudolph KS, Axe MJ, Snyder-Mackler L. Dynamic stability after ACL injury: who can hop? *Knee Surg Sports Traumatol Arthrosc* 2000;**8**:262–9.
- Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. *Clin Orthop Relat Res* 1985;**198**:43–9.
- Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;**18**:979–92.
- Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;**118**:622–9.
- Reid A, Birmingham TB, Stratford PW, et al. Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Phys Ther* 2007;**87**:337–49.
- Newton RU, Gerber A, Nimphius S, et al. Determination of functional strength imbalance of the lower extremities. *J Strength Cond Res* 2006;**20**:971–7.
- Urbach D, Awiszus F. Impaired ability of voluntary quadriceps activation bilaterally interferes with function testing after knee injuries: a twitch interpolation study. *Int J Sports Med* 2002;**23**:231–6.
- Rockborn P, Gillquist J. Outcome of arthroscopic meniscectomy: a 13-year physical and radiographic follow-up of 43 patients under 23 years of age. *Acta Orthop Scand* 1995;**66**:113–17.
- Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil* 2000;**81**(Suppl 2):S15–20.
- Fuller CW, Hawkins RD. Developing a health surveillance strategy for professional footballers in compliance with UK health and safety legislation. *Br J Sports Med* 1997;**31**:148–9.
- Schunemann HJ, Guyatt GH. Commentary: goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* 2005;**40**:593–7.
- Irrgang JJ, Anderson AF, Boland AL, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2006;**34**:1567–73.
- Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;**7**:541–6.
- de Vet HC, Terwee CB, Ostelo RW, et al. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;**4**:54.
- Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;**61**:102–9.
- Frost MH, Reeve BB, Liepa AM, et al. Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health* 2007;**10**(Suppl 2):S94–105.
- Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Stat Med* 2002;**21**:3431–46.
- Batterham AM, George KP. Reliability in evidence-based clinical practice: a primer for allied health professionals. *Phys Ther Sport* 2003;**4**:122–8.
- Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*, 3rd edn. Oxford: Oxford University Press, 2003.
- Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;**26**:217–38.
- Keller A, Hellesnes J, Brox JI. Reliability of the isokinetic trunk extensor test, Biering-Sorensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. *Spine* 2001;**26**:771–7.

## Review

46. **Weir JP.** Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005;**19**:231–40.
47. **Nunnally JC,** Bernstein IH. *Psychometric theory*, 3rd edn. New York: McGraw-Hill, 1994.
48. **Hays RD,** Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993;**2**:441–9.
49. **Bolgla LA,** Keskula DR. Reliability of lower extremity functional performance tests. *J Orthop Sports Phys Ther* 1997;**26**:138–42.
50. **Schmitt JS,** Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol* 2004;**57**:1008–18.
51. **Bland JM,** Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307–10.
52. **Marx RG,** Jones EC, Allen AA, *et al.* Reliability, validity, and responsiveness of four knee outcome scales for athletic patients. *J Bone Joint Surg Am* 2001;**83A**:1459–69.
53. **Jerrosch-Herold C.** An evidence-based approach to choosing outcome measures: a checklist for the critical appraisal of validity, reliability and responsiveness studies. *Br J Occup Ther* 2005;**68**:347–53.
54. **Switzer GE,** Wisniewski SR, Belle SH, *et al.* Selecting, developing, and evaluating research instruments. *Soc Psychiatry Psychiatr Epidemiol* 1999;**34**:399–409.
55. **Turner RR,** Quittner AL, Parasuraman BM, *et al.* Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. Patient-reported outcomes: instrument development and selection issues. *Value Health* 2007;**10**(Suppl 2):S86–93.
56. **Brosseau L,** Balmer S, Tousignant M, *et al.* Intra- and intertester reliability and criterion validity of the parallelogram and universal goniometers for measuring maximum active knee flexion and extension of patients with knee restrictions. *Arch Phys Med Rehabil* 2001;**82**:396–402.
57. **Irrgang JJ,** Snyder-Mackler L, Wainner RS, *et al.* Development of a patient-reported measure of function of the knee. *J Bone Joint Surg Am* 1998;**80A**:1132–45.
58. **Goodwin LD.** Changing conceptions of measurement validity. *J Nurs Educ* 1997;**36**:102–7.
59. **Beaton DE,** Bombardier C, Katz JN, *et al.* A taxonomy for responsiveness. *J Clin Epidemiol* 2001;**54**:1204–17.
60. **Stratford PW,** Riddle DL. Assessing sensitivity to change: choosing the appropriate change coefficient. *Health Qual Life Outcomes* 2005;**3**:23.
61. **Husted JA,** Cook RJ, Farewell VT, *et al.* Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* 2000;**53**:459–68.
62. **Terwee CB,** Dekker FW, Wiersinga WM, *et al.* On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res* 2003;**12**:349–62.
63. **Holtby R,** Razmjou H. Measurement properties of the Western Ontario rotator cuff outcome measure: a preliminary report. *J Shoulder Elbow Surg* 2005;**14**:506–10.
64. **Roos EM,** Roos HP, Lohmander LS, *et al.* Knee injury and Osteoarthritis Outcome Score (KOOS): development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;**28**:88–96.
65. **Cortina JM.** What is coefficient alpha: an examination of theory and applications. *J Appl Psychol* 1993;**78**:98–104.
66. **Batterham AM,** Atkinson G. How big does my sample need to be? A primer on the murky world of sample size estimation. *Phys Ther Sport* 2005;**6**:153–63.
67. **Garratt AM,** Brealey S, Gillespie WJ, DAMASK Trial Team. Patient-assessed health instruments for the knee: a structured review. *Rheumatology* 2004;**43**:1414–23.



## Selecting outcome measures in sports medicine: a guide for practitioners using the example of anterior cruciate ligament rehabilitation

N P Bent, C C Wright, A B Rushton, et al.

*Br J Sports Med* 2009 43: 1006-1012 originally published online February 17, 2009  
doi: 10.1136/bjasm.2009.057356

---

Updated information and services can be found at:  
<http://bjsm.bmj.com/content/43/13/1006.full.html>

---

*These include:*

### References

This article cites 63 articles, 10 of which can be accessed free at:  
<http://bjsm.bmj.com/content/43/13/1006.full.html#ref-list-1>

### Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

---

### Notes

---

To request permissions go to:  
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:  
<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:  
<http://group.bmj.com/subscribe/>