APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

Population studied		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
Reliability		
Reliability Description	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same health-related instrument (internal consistency), over time (test retest) by different persons on the same occasion (interrater) or by the same persons (i.e., raters or responders) on different occasions (intra-rater)	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa(Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Measurement Error	The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation); 2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their

		mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Validity		
Content Validity	The degree to which the content of a health-related instrument is an adequate reflection of the construct to be measured	A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found. 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents; their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions 5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions 6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes
Construct Validity	The degree to which the scores of a health-related instrument are consistent with hypotheses (for instance with regard to internal	When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct

	relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the health-related instrument validly measures the construct to be measured	validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.
Structural Validity/Hypothesis Testing	The degree to which the scores of a health-related instrument are an adequate reflection of the dimensionality of the construct to be measured	Extract test names, prespecified expected relationship and correlations observed.
Structural validity - discriminative	discriminative analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and known to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.

Criterion validity	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations or other as reported.
Responsiveness/Clinical Change		
Responsiveness	The ability of a health-related instrument to detect change over time in the construct to be measured	Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)
Interpretability		
Interpretability	The degree to which one can assign qualitative meaning that is, clinical or commonly understood connotations to an instrument's quantitative scores or change in scores.	

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year:_____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
Content/structural validity	
Internal consistency	
Content Validity	

Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	
Responsiveness	

Minimally Clinical Important Difference	

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group)_____
Author(s) (Study Author(s) _____
Year (Year of publication)_____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?
   2
   1
   0
2. Were appropriate inclusion/exclusion criteria defined? *
   2
   1
   0
3. Were specific clinical measurement questions/hypotheses identified?
   2
   1
   0
4. Was an appropriate scope of measurement properties considered?
   2
   1
   0
5. Was an appropriate sample size used?
   2
   1
   0
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)
   2
   1
   0

7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?

2
1
0

8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?

2
1
0

9. Were analyses conducted for each specific hypothesis or purpose?

2
1
0

10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?

2
1
0

11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?

2
1
0

12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?

2
1
0

Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks[25,26,27,28,29,30]	Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg. Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test. One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand. Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec). Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec). Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.

A Physiotherapy Test Package[33,34,35,36]	**PILE Tests:** "The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 5.9 kg for men. A 'lifting movement' involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the 'adjusted weight'". **2x20m WWB:** "Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s".
BTEWS II[31]	"The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level".
FIT - HaNSA[32]	"The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately "grab, lift, move and place" three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye-down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf