COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument

user manual

Version 1.0 dated January 2021

Lidwine B Mokkink
Maarten Boers
Cees van der Vleuten
Donald L Patrick
Jordi Alonso
Lex M Bouter
Henrica CW de Vet
Caroline B Terwee

Contact
LB Mokkink, PhD
Amsterdam UMC, Vrije Universiteit Amsterdam,
Department of Epidemiology and Data Science
Amsterdam Public Health research institute
De Boelelaan 1117, 1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl
E-mail: w.mokkink@amsterdamumc.nl
The development of the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error was part of the VENI programme with project number 91617098, funded by ZonMw (The Netherlands Organisation for Health Research and Development).
# Table of Content

Foreword 5

1. Background information 6
   1.1 COSMIN initiative and steering committee 6
   1.2 How to cite this manual 7
   1.3 Development of the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error 7
   1.4 Definitions of reliability and measurement error 7
   1.5 Focus of the COSMIN Risk of Bias tool 8
   1.6 The structure of the COSMIN Risk of Bias tool 10
   1.7 The “worst-score-counts” method 10
   1.8 Relevance of the research question 11
   1.9 Using the COSMIN Risk of Bias tool in a systematic review 11
   1.10 Expertise required for using the tool 12
   1.11 Using the COSMIN Risk of Bias tool to assess studies on PROMs or ObsROMs 12
   1.12 A Risk of Bias tool is not a study design checklist, nor a reporting guideline 13

2. Part A. Understanding how a study informs us about the reliability and measurement error of an outcome measurement instrument 14
   2.1 Components of outcome measurement instruments 14
   2.2 Extracting the elements of a comprehensive research question 20
   2.3 Example of how to use Part A of the COSMIN Risk of Bias tool to assess the quality of a study by Skeie et al. (2015) 27

3. Part B. Assessing the risk of bias of a study on reliability or measurement error 31
   3.1 Elaboration on standards for studies on reliability 33
   3.2 Elaboration on standards for studies on measurement error 40
   3.3 Example of how to use Part B of the COSMIN Risk of Bias tool to assess the quality of a study by Skeie et al. (2015) 45

4. Using the COSMIN Risk of Bias tool in a systematic review of outcome measurement instruments 47
   4.1 The eleven-step procedure for conducting a systematic review of ClinROMs, PerFOMs, or laboratory values 50
Appendix 1. Data Extraction table of relevant information for each included study in a systematic review. 60
Appendix 2. Risk of Bias ratings per standard per study 62
Appendix 3. Example of a Flow-chart 63
Appendix 4. Example of reporting table on characteristics of the included measurement instruments. 64
Appendix 5. Example of reporting table on characteristics of the study populations. 65
Appendix 6. Overview Table of quality and results of studies on reliability and measurement error. 66
Appendix 7. Summary of Findings Tables for Reliability and Measurement error. 67
References 68
Foreword

The COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error was developed to transparently and systematically assess the methodological quality of studies on reliability and measurement error of all types of outcome measurement instruments. It is an extended version of the COSMIN Risk of Bias checklist for the boxes reliability and measurement error for PROMs (1). It was developed for clinician-reported outcome measures (ClinROMs) (including e.g. readings based on imaging modalities and ratings based on observations), performance-based outcome measurement instruments (PerFOMs), or biomarkers – also called laboratory values (2, 3). These measurement instruments are more complex than PROMs, as not only patients are involved, but also professionals, and sometimes (complex) devices. Specifically in studies on reliability and measurement error these additional sources of variation complicate the design of these studies and may influence their quality.

As different sources of variation can play a role, different studies can be conducted to assess the reliability or measurement error of an outcome measurement instrument. To assess the quality of such a study, one should understand (1) how the results of a published study on reliability or measurement error inform us about the reliability and measurement error of the outcome measurement instrument under study, and (2) whether we can trust the result found in the study by assessing the risk of bias of the study. These two steps are reflected in the new COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments (4).

The quality assessment of a study on reliability or measurement error can be conducted in the context of a systematic review of outcome measurement instruments. In such a review all measurement properties are considered, the quality of the each study is assessed, the results of the studies are extracted, and per measurement property an overall conclusion is drawn about the quality of the instrument based on all available evidence for each measurement instrument. Subsequently, the quality of the evidence is graded, taking the number, quality, and (consistency of) results of the studies into account. A recommendation for the most suitable instrument is made, based on quality, feasibility and interpretability of each instrument.

As this is not an easy task to perform, we encourage to use systematic and transparent methods when conducting such systematic reviews. We developed the COSMIN methodology for conducting systematic reviews of PROMS (5), including the COSMIN Risk of Bias checklist (1, 6). When conducting a systematic review of other types of outcome measurement instruments, such as ClinROMs, PerFOMs, or laboratory values, this newly developed COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error can be incorporated into the COSMIN methodology. In this manual we will explain how this new tool should be used.
1. Background information

1.1 COSMIN initiative and steering committee

The COSMIN initiative aims to improve the selection of health measurement instruments both in research and clinical practice by developing tools for selecting the most suitable instrument for a given situation. COSMIN is an international initiative consisting of a multidisciplinary team of researchers with expertise in epidemiology, psychometrics, and qualitative research, and in the development and evaluation of outcome measurement instruments in the field of health care, as well as in performing systematic reviews of outcome measurement instruments.

This tool was developed in a Delphi study (4). The steering committee of this study consisted of:

Lidwine B Mokkink
Maarten Boers
Cees van der Vleuten
Donald L Patrick
Jordi Alonso
Lex M Bouter
Henrica CW de Vet
Caroline B Terwee

We are very grateful to all the panelists of this study, who provided us with many helpful and critical comments and arguments (in alphabetical order): M.A. D’Agostino, Dorcas Beaton, Sophie van Belle, Sandra Beurskens, Kristie Bjornson, Jan Boehnke, Patrick Bossuyt, Don Bushnell, Stefan Cano, Saskia le Cessie, Alessandro Chiarotto, Mike Clark, Jon Deeks, Iris Eekhout, Jim Farnsworth II, Oke Gerke, Sabine Goldhahn, Robert M. Gow, Philip Griffths, Cristian Gugi, Jean-Benoit Hardouin, Desirée van der Heijde, I-Chan Huang, Ellen Janssen, Brian Jolly, Lars Konge, Jan Kottner, Brittany Lapin, Hanneke van der Lee, Mariska Leeflang, Nancy Mayo, Sue Mallett, Joy C. MacDermid, Geert Molenberghs, Holger Muehlan, Koen Neijenhuijs, Raymond Ostelo, Laura Quinn, Dennis Revicki, Jussi Repo, Johannes B. Reitsma, Anne W. Rutjes, Mohsen Sadatsafavi, David Streiner, Matthew Stephenson, Berend Terluin, Zyphanie Tyack, Werner Vach, Gemma Vilagut Saiz, Marc K. Walton, Matthijs Warrens, and Daniel Yee Tak Fong.
1.2 How to cite this manual

This manual accompanies the tool developed in the Delphi study. Please, refer to the article when using the manual of the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error.

LB Mokkink, M Boers, CPM van der Vleuten, LM Bouter, J Alonso, DL Patrick, HCW de Vet, CB Terwee. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Medical Research Methodology. 2020;20(293).

1.3 Development of the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error

This COSMIN tool was developed in a Delphi study, containing three rounds. For more information about the methods of this study, we refer to Mokkink et al. 2020. In this Delphi study we reached consensus on how to formulate a comprehensive research question for studies on reliability and measurement error, on components of outcome measurement instruments (which are the potential sources of variation relevant in studies on reliability and measurement error), and on standards to assess the quality of a study on reliability and measurement error of ClinROMs, PerFOMs, or laboratory values. Based on those results, we developed the COSMIN Risk of Bias tool which comprises two parts: 1) seven elements that make up a comprehensive research question of the study, which informs us on how the reliability and measurement error of the outcome measurement instrument was studied, and 2) standards on design requirements and preferred statistical methods of studies on reliability and measurement error, which can be used to assess the quality of the study.

1.4 Definitions of reliability and measurement error

Reliability and measurement error are important measurement properties of outcome measurement instruments. Reliability and measurement error are determined based on the same study design and data collection, but with different statistical methods. These measurement properties are therefore related, but distinct.

**Reliability** is defined as the proportion of the total variance in the measurement which is due to true differences between patients (7). It refers to what extend an instrument is able to distinguish between patients; a reliability study investigates the extent to which different sources of variation influence the measurement. This gives direction for how to improve the measurement, for example by standardization or restriction of the source of variation. Reliability can be calculated with an Intra-class Correlation
Coefficient (ICC), a Generalizability Coefficient or with a kappa. Reliability parameters are expressed as a proportion and lies between 0 and 1.

Measurement error is defined as the systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured (7). It refers to how close the scores of repeated measurements in stable patients are; such studies investigate the absolute deviation of the scores or the amount of error of repeated measurements in stable patients.

In case of categorical outcomes it is also called ‘agreement’. For continuous outcomes measurement error is expressed in the measurement units of the measurement instrument with a Standard Error of Measurement (SEM) or Limits of Agreement (LoA). For categorical outcomes agreement is expressed as percentage total agreement or percentages specific (e.g. positive and negative) agreement.

1.5 Focus of the COSMIN Risk of Bias tool

We focus on outcome measurement instruments, defined as instruments used to monitor the health status of (a group of) people over time, for example in a clinical trial or in clinical practice.

Several types of measurement instruments exist, such as patient-reported outcome measure (PROM); observer-reported outcome measures (ObsROMs; i.e. proxy measures); clinician-reported outcome measurement instruments (ClinROMs) (including e.g. readings based on imaging modalities and ratings based on observations); performance-based outcome measurement instruments (PerFOMs); and biomarker outcomes – also called laboratory values (2).

The COSMIN Risk of Bias tool to assess reliability and measurement error is specifically developed for ClinROMs, PerFOMs, and laboratory values (see Table 1 for examples). These outcome measurement instruments typically require involvement of one or more professionals to operate equipment or tools, to give instructions to the patient (e.g. to perform a task or action) or to come to a score through their clinical expertise (e.g. after observing a patient or an image). An outcome measurement instrument comprises the whole measurement procedure to come to a score, including issues such as materials, communication (e.g. instructions and motivating patients in case of performance-based test), clinical judgment, performing a task. All issues relevant for reliable and valid measurement should be described in the measurement protocol of an outcome measurement instrument.
**Table 1. Examples of ClinROMs, PerFOMs, and laboratory values**

<table>
<thead>
<tr>
<th>Clinician-reported outcome measurement instruments (ClinROMs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician-reported rating of the severity of a disease or condition. For example, the Hamilton Anxiety Rating Scale to assess the severity of anxiety symptoms comprises 14 items that are scored by a clinician (8).</td>
</tr>
<tr>
<td>A Global Assessment of the severity of a condition scored e.g. on a single-item Visual Analogue Scale by a health-care professional.</td>
</tr>
<tr>
<td>Result of clinical examination of (patho)physiology, such as blood pressure or a count of swollen joints.</td>
</tr>
<tr>
<td>Clinical reading of device-based results (often imaging), such power Doppler ultrasonography to assess s cardiac structure, function and hemodynamics (echocardiography) (9), or MRI used to evaluate cartilage defect size, depth, and subchondral bone in order to assess chondral and osteochondral lesions at the knee (10).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance-based outcome measurement instrument (PerFOMs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A performance-based walking test (e.g. the timed 25-foot walk test (11)), in which a professional instructs a patient to walk 25 feet at his own comfortable pace with or without a walking aid. Time needed to cover 25 feet is measured by the professional.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Laboratory value or biomarker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory value such as HbA1c (glycated haemoglobin) measured by the turbidimetric inhibition immunoassay (TINIA) (12).</td>
</tr>
</tbody>
</table>

**Different versions or operationalizations of outcome measurement instruments**

To measure a specific construct, different versions of a measurement instrument may exist. For example, the Doloplus is a clinical assessment tool to measure behavioural pain assessment in cognitively impaired patients, and is administered e.g. by the attending nurse. The original Doloplus-1 contained 15 items, while the Doloplus-2 contains 10 items (13). A measurement instrument (i.e. the measurement protocol) can be operationalized in many different ways, and each operationalization could be considered a different version. For example, the specific equipment used to measure the range of motion (ROM) can differ, e.g., a simple universal goniometer (14) or an electromagnetic 3-dimensional tracking system (15). The location to be measured can differ, e.g., the neck (14) or the shoulder (16). The background of the professional involved can differ, e.g., a rheumatologist or a radiologist who conducts the measurement, and these raters may have had different levels of training (17).

In principle, we consider each version of an outcome measurement instrument or each different operationalization of the measurement protocol as a separate measurement instrument, until evidence is provided (e.g. testing of measurement invariance, or reliability) that the versions perform similarly.
1.6 The structure of the COSMIN Risk of Bias tool

The COSMIN Risk of Bias tool comprises two parts. Part A helps to understand how the results of a published study inform us about the reliability or measurement error of the outcome measurement instruments under study. Part B helps to assess whether we can trust the result obtained in the study by assessing the risk of bias of the study.

Part A
For a good understanding of how the results of a study informs us about the reliability and measurement error of the instrument, a good understanding of the design of the study and its corresponding comprehensive research question is needed. In part A we describe the seven elements that we recommend to be extracted, and that together can be used to construct a comprehensive research question for each analysis. In addition, Part A of the tool contains an overview of the components of outcome measurement instruments. These component are the potential sources of variation that can either be studied (i.e. varied across the repeated measurements), or are kept or assumed to be stable (i.e. standardized).

Part B.
Next, we developed two boxes with standards for studies on reliability and for studies on measurement error, respectively. As in the COSMIN Risk of Bias checklist for PROMs (1), standards refer to design requirements and preferred statistical methods of studies on measurement properties. For example, ‘reliability and measurement error should be assessed in patients that are assumed to be stable’; or ‘measurement error should be assessed with the standard error of measurement or with the limits of agreement’. The standards are stated as questions: e.g. ‘were patients stable in the interim period on the construct to be measured?’

We refer to ‘preferred’ statistical methods. We mean by ‘preferred’ that these statistical methods are appropriate to use when evaluating reliability or measurement error of outcome measurement instruments, and are commonly used. Other methods may be appropriate to use as well (for example bi-factor models or Multi-Trait Multi-Method (MTMM) analyses, or newly developed methods). It is not our intention to comprehensively describe all possible statistical methods, rather to describe the adequate methods that are commonly used in the literature. It is up to the user of the COSMIN tool how studies using these less commonly used methods are assessed.

1.7 The “worst-score-counts” principle

Each standard in a box is scored on the four-point scale, i.e. ‘very good’, ‘adequate’, ‘doubtful’, and ‘inadequate’, see chapter 3 for more information. Similar as in the COSMIN Risk of Bias checklist for PROMs (1), we use the worst-score-counts method (18) to come to a rating for the quality of the study on reliability or measurement error.
1.8 Relevance of the research question

While many different research questions concerning the reliability or measurement error of an outcome measurement instrument can be investigated, the relevance of a study is not under question when using this tool. The relevance of a study refers to different aspects.

- Choice of the potential source(s) of variation that has been varied over the repeated measurements.
- Choice of the target population of patients and professionals (when applicable) of the study.
- Choice of how the measurement protocol was executed, when applicable.
- Choice of evaluating the specific measurement property, either reliability or measurement error. Often only reliability is reported, while the measurement error can be calculated using the same data.

When using this COSMIN Risk of Bias tool, these aspects will be extracted from the design of the study (in part A). However, no judgement will be given about the appropriateness of the choices made. The choices made in the research question and study design by the researchers determine the interpretation and generalizability of the results.

1.9 Using the COSMIN Risk of Bias tool in a systematic review

The COSMIN Risk of Bias tool is developed to assess the quality of a published study. One application of the COSMIN Risk of Bias tool is to assess the quality of studies when conducting a systematic review on measurement instruments. COSMIN developed a systematic methodology for conducting systematic reviews of PROMs (5). It consists of a 10 step procedure, in which the COSMIN Risk of Bias checklist (1)(containing standards for all nine measurement properties) can be applied to the studies to assess the quality of each study. To use the COSMIN methodology for conducting systematic reviews of other types of instruments – that is: other than PROMs – we advise to replace the boxes 6 (Reliability) and 7 (Measurement error) with the COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instruments. More information about how to conduct a systematic review using the new COSMIN Risk of Bias tool can be found in chapter 4.
1.10 Expertise required for using the tool

To assess the quality of a study on reliability and measurement error, i.e. for use in a systematic review on the quality of outcome measurement is quite complex and time consuming, and it requires expertise within the research team on several aspects. We recommend that at least one of the team members should have expertise on the construct to be measured, e.g. to understand what appropriate time intervals are between repeated measurements; on the measurement instruments, e.g. to understand what concomitant sources of variation could be (and these should be restricted or standardized – see element 2 in Part A); on the patient population, e.g. to understand whether patients were stable between repeated measurements or whether subgroups of patients can be considered in one study. A clinical expert might combine these expertises. A methodological expert should be part of the team member with expertise on the theory of reliability and measurement error, e.g. to understand whether the design is appropriately analyzed (e.g. standards 7).

1.11 Using the COSMIN Risk of Bias tool to assess studies on PROMs or ObsROMs

This new COSMIN Risk of Bias tool is developed specifically for ClinROMs, PerFOMs, and laboratory values. However, it can also be used to assess the quality of studies on reliability or measurement error of PROMs or observer-reported outcome measures (ObsROMs; i.e. observations made, appraised, and recorded by a person other than the patient who does not require specialized professional training (2), e.g. proxy measures). However, for these two types of instruments the tool may seem unnecessarily complex. The first step in the tool (i.e. understanding how the results inform us on the quality of the measurement instrument under study) is often obvious, as the aim of reliability studies of PROMs and ObsROMs is most often to assess test-retest reliability or measurement error of the whole measurement instrument (as these measurement instruments can only be taken in one go, and the only potential source of variance is occasion). The second step in the tool (assessing the quality of the study using the standards) will lead to the same rating compared to using the standards of the Risk of Bias checklist for PROMs. The standards on design requirements in both tools are partly the same. However, the new types of outcome measurement instruments for which we adapted the COSMIN checklist (i.e. ClinROMs, PerFOMs and laboratory values), require additional standards, which are not usually applicable for PROMs and ObsROMs. (If it is applicable in a specific study, it could be rated using the ‘other flaws’ standard in the Risk of Bias checklist for PROMs). The response options for standards on preferred statistical methods in the new tool are somewhat differently formulated, but will lead to the same rating as the PROM Risk of Bias checklist.
1.12 A Risk of Bias tool is not a study design checklist, nor a reporting guideline

This COSMIN Risk of Bias tool is developed to assess the quality (i.e. risk of bias) of a published study on reliability or measurement error. This tool is not developed as a design checklist or a reporting guideline. When designing or reporting a study on reliability or measurement error additional items are relevant to consider or report. For example, the sample size of patient samples and number of raters or repeated measurements are important in the design of a study, and when reporting specific results such as the variance components, 95% confidence intervals around ICCs, marginal when reporting kappa’s, or additional assumptions are required.
Part A. Understanding how a study informs us about the reliability and measurement error of an outcome measurement instrument.

In general, the design of a study on reliability and measurement error is about repeated measurement in stable patients. Each measurement is accompanied by some error. This error is caused by sources of variation, such as the equipment used, the professionals involved, and other components of measurement instruments. For example, the score on an instrument can be influenced by how the rater motivates the patient, how the machine was set up, or by the occasion (e.g. first and second occasion, day of the week, time of the day).

In chapter 2.1 we systematically describe all components of outcome measurement instruments, which are the potential sources of variation of an outcome measurement instrument. Many different sources of variation can affect the measurement, and each of them can be studied using a different study designs. Each study design answers a different research question, and each research question gives specific information about the quality of the measurement instrument. To understand how a study can inform us about the quality of an outcome measurement instrument we describe in chapter 2.2 seven elements of a comprehensive research question. Part A of the tool contains the overviews of the components of outcome measurement instruments (for outcome measurement instruments that does not involve biological sampling, and those that involve biological sampling, respectively), and the seven elements of a comprehensive research question.

In chapter 2.3 we provide an example in which we show how to use Part A of the tool, by applying it to a paper by Skeie (19). In chapter 2.2 we will use this example, too (among other examples).

2.1 Components of outcome measurement instruments

All measurement instruments consist of components, such as equipment and preparatory actions. We developed two taxonomies of components of outcome measurement instruments, one for outcome measurement instruments that do not involve biological sampling (i.e. ClinROMs and PerFOMs) (see Table 2), and one for those that do (i.e. the laboratory values, such as blood or urine tests, tissue biopsy) (see Table 3).
Table 2. Components of outcome measurement instruments that do not involve biological sampling

<table>
<thead>
<tr>
<th>Component</th>
<th>Elaboration</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment</td>
<td>All equipment necessary in the preparation, the administration, and the assignment of scores of the outcome measurement instrument</td>
<td>Questionnaire forms, computers, tablet, pen and paper; stair steps of a specific height; device or tools (such as stopwatch, probe, tube); ultrasound machine, ultrasound gels, MRI scanner; software.</td>
</tr>
</tbody>
</table>
| Preparatory actions preceding raw data collection by professionals, patients, and others (if applicable) | 1. General preparatory actions, such as required expertise or training for professionals to prepare, administer, store or assign the scores  
2. Specific preparatory actions for each measurement, such as  
   - preparations of equipment, environment, storage by professionals\(^a\)  
   - preparations of the patient\(^b\) by the professional | Training, education or experience required, certification.  
Preparation of equipment: calibration of device/equipment, adjust settings of the machine.  
Preparation of the environment: light conditions, room temperature, humidity, specific length of a walking track.  
Preparation for storage: design database and logbook  
Provide general and preparatory instructions for the patients, such as explaining the tasks/action that need to be performed including time schedule, safety issues and side effects; instructions on diet (e.g. use of caffeine), clothing (e.g. comfortable shoes, no jewelry, glasses or devices), performance during tests (e.g. perform a task as usual; try to walk as fast as you can; lie as calm as possible); set some training or perform a familiarization session.  
Attaching electrodes to the body, injection with radioactive substance or contrast dye, positioning the patient, applying ultrasound gel. |
<table>
<thead>
<tr>
<th>Component</th>
<th>Elaboration</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Preparations undertaken by the patients</td>
<td>Listen to and understanding the instructions provided; adherence to the preparatory instructions such as fasting, resting, taking medication, bowel preparation, exercising, shaving.</td>
</tr>
<tr>
<td>Collection of raw data</td>
<td>All actions undertaken by patient and professional(s) to collect the data, before any data processing</td>
<td>The patient completing questions at home, or at the hospital; or performing the tasks; the rater observing or timing the performance; switching the imaging device on and off; positioning and moving the ultrasound probe.</td>
</tr>
<tr>
<td>Data processing and storage</td>
<td>All actions undertaken on the raw data to store it in a usable (electronic) form for later data manipulation (such as score assignment or statistical analysis)</td>
<td>The digitally converted signal of a specific body MRI scan which is temporarily stored in the K-space, is sent to an image processor where a mathematical formula (i.e. Fourier transformation) is applied, leading to an image which is displayed on a monitor and saved on a computer; Other examples: answers of question items are recorded on e.g. paper forms and stored or Likert scale format response options are converted into a 0-4 score and directly entered in a computer database. Performance of data quality checks e.g. double entry or validation checks on the stored/entered data.</td>
</tr>
<tr>
<td>Assignment of the score(s)</td>
<td>Methods used to convert processed data into a score that constitutes the outcome measurement instrument.</td>
<td>A calculation of a mathematical formula or the application of a scorings algorithm (e.g. a set of rules to be followed) to the processed data; a clinician selects the specific images and judges the severity and quantity of e.g. lesions on the set of images or compares it to a reference; scores adjusted for e.g. missing data or patients using devices such as mobility aids.</td>
</tr>
</tbody>
</table>

*a Professionals are those who are involved in the preparation or the performance of the measurement, in the data processing, or in the assignment of the score; this may be done by one and the same person, or by different persons. **In the COSMIN methodology we use the word ‘patient.’ However, sometimes the target population is not patients, but e.g. healthy individuals, caregivers, clinicians, or body structures (e.g. joints, or lesions). In these cases, the word patient should be read as e.g. healthy volunteer, clinician, or the relevant body structure. **The score can be further used or interpreted, by converting a score to another scale, metric or classification. For example, a continuous score is classified into an ordinal score (e.g. mild/moderate/severe), a score is dichotomized into below or above a normal value, patients are classified as responder to the intervention (e.g. when their change is larger than the Minimal Important Change (MIC) value).
Table 3. Components of outcome measurement instruments that involve biological sampling

<table>
<thead>
<tr>
<th>Component</th>
<th>Elaboration</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equipment</td>
<td>All equipment used in the preparation, the administration, and the determination of the values of the outcome measurement instrument</td>
<td>Collection tools, such as vena puncture set, biopsy tool; material containers, such as for blood plasma (EDTA of heparin tube), for tissue (container for frozen specimens for immunofluorescence, jar filled with formalin), for urine collection (sterile, screw-top container), for standard microscopic tissue evaluation (fluid or tissue for culture (sterile jar)); laboratory equipment such as centrifuges, cabinets, and chromatography systems, computers, software.</td>
</tr>
<tr>
<td>Preparatory actions preceding sample collection by professionals, patients, and others (if applicable)</td>
<td>1. General preparatory actions, such as required expertise or training for professionals to prepare, administer, store and determine the value</td>
<td>Training, education or experience required, certification.</td>
</tr>
<tr>
<td></td>
<td>2. Specific preparatory actions for each measurement, such as</td>
<td>Preparation of equipment: calibration of device/equipment, adjust settings of the machine. Preparation of the environment: light conditions, room temperature, humidity. Preparation of storage: set-up all equipment for storage. Provide general and preparatory instructions to the patients, such as explaining the measurement procedure including safety issues and side effects; instructions on diet; insertion and withdrawal of a catheter into a blood vessel.</td>
</tr>
<tr>
<td>Component</td>
<td>Elaboration</td>
<td>Examples</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Collection of biological sample</td>
<td>All actions undertaken to collect the biological sample, before any sample processing</td>
<td>Taking a blood sample or tissue biopsy, collection of a sample of urine 'mid-stream' in a container.</td>
</tr>
<tr>
<td>Biological sampling processing and storage</td>
<td>All actions undertaken to be able to preserve, transport, and store the biological sample for determination; and, if applicable, further actions undertaken on the stored sample to be able to conduct the determination of the biological sample</td>
<td>Initial reaction of material to reagent in container (e.g. anticoagulation by heparin). Blood is decomposed (by gravity) into plasma and blood cells, and stored at a specific temperature. Tissue is snap frozen by immersion in liquid nitrogen, or fixed in formalin embedded in/processed to paraffin for long-term storage. Blood is collected in a tube containing an aqueous solution tetra-sodium salt of ethylene-diamine-tetra-acetic acid (EDTA) and mixed with air to lyse the erythrocytes and convert hemoglobin to oxyhemoglobin. Cut sections or prepare a smear on a slide, tissues are stained by immunofluorescent markers specific for certain surface antigens. Screw the lid of the urine container shut, put in a sealed plastic bag and store it in the fridge at around 4 degrees Celsius, for max. 24 hours.</td>
</tr>
<tr>
<td>Determination of the value of the biological sample</td>
<td>Methods used for counting or quantifying the amount of the substance or entity of interest</td>
<td>The absorbance of oxyhemoglobin at 540 nm through spectrophotometry quantifies the hemoglobin concentration in the sample. The presence of the marker on the cell surface is detected and quantified by fluorescence signal intensity. Rater observes each slide and counts positive cells in an area. A calculation or the application of a mathematical formula to the prepared sample.</td>
</tr>
</tbody>
</table>
Professionals are those who are involved in the preparation or the performance of the measurement, in the data processing, or in the assignment of the score; this may be done by one and the same person, or by different persons. In the COSMIN methodology we use the word ‘patient.’ However, sometimes the target population is not patients, but e.g. healthy individuals, caregivers, clinicians, or body structures (e.g. joints, or lesions). In these cases, the word patient should be read as e.g. healthy volunteer, clinician, or relevant body structure.

The value can be further processed into a clinical score, if applicable, by a linear or semi-quantitative conversion. For example, a continuous score is classified into an ordinal score (e.g. mild/moderate/severe), a scores is dichotomized into below or above a normal value, patients are classified as responder on treatment (e.g. when their change is larger than the Minimal Important Change (MIC) value). As no noise will occur from this conversion, this is not a potential source of variance, but rather an interpretation of the value. Therefore we do not include this phase in the components for outcome measurement instruments that involve biological materials.
2.2 Extracting the elements of a comprehensive research question

Before we can comprehensively assess the information in a study on the reliability or measurement error of an instrument, we need to fully understand the design of the study and reformulate the research question into what we call a ‘comprehensive research question’. Often the published research question is not specific enough to rate the adequacy of the study design. For example, if the stated aim of their study is to assess inter-rater reliability of an instrument, it is clear that *raters* will be varied. However, without further information it is not clear whether the interest is in the inter-rater reliability of the whole measurement procedure (e.g. by different clinicians), or only in the reliability of a part of the measurement procedure (e.g. only the assignment of the score based on an image).

To get a complete picture, we recommend to extract seven elements from the publication that together can form the ‘comprehensive research question’ (see Table 4). Note that one article can contain multiple questions, each requiring an extraction of the seven elements.

Table 4. Elements of a comprehensive research question.

<table>
<thead>
<tr>
<th>1</th>
<th>the name of the outcome measurement instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>the version of the outcome measurement instrument or way of operationalization of the measurement protocol</td>
</tr>
<tr>
<td>3</td>
<td>the construct measured by the measurement instrument</td>
</tr>
<tr>
<td>4</td>
<td>a specification whether one is interested in a reliability parameter (i.e. a relative parameter such as for continuous outcomes an ICC, Generalizability coefficient ( \varphi ), or Kappa ( \kappa )) or a parameter of measurement error (i.e. an absolute parameter expressed in the unit of measurement e.g. SEM, LoA or SDC; or for categorical outcomes expressed as agreement or misclassification, e.g. the percentage specific agreement).</td>
</tr>
<tr>
<td>5</td>
<td>a specification of the components of the measurement instrument that will be repeated (especially when only part of the measurement instrument is repeated, e.g. only assignment of the score based on the same images)</td>
</tr>
<tr>
<td>6</td>
<td>a specification of the source(s) of variation that will be varied (e.g. time or occasion, the (level of expertise of) professionals, the machines, or other components of the measurement)</td>
</tr>
<tr>
<td>7</td>
<td>a specification of the patient population studied</td>
</tr>
</tbody>
</table>

**ICC** = Intraclass correlation coefficient; **SEM** = standard error of measurement; **LoA** = Limits of Agreement; **SDC** = smallest detectable change.
Elaboration on the elements of a comprehensive research question

Element 1. The name of the outcome measurement instrument

The name of the instrument should be exactly specified. Sometimes, this is readily apparent, e.g. the 6 minute Walking test (6MWT) or the Nine Hole Peg Test (NHPT). In some cases, a measurement protocol involves multiple measurement instruments (e.g. the Multiple Sclerosis Functional Composite (MSFC) includes the Timed 25-Foot Walk test, the Nine Hole Peg Test, and the Paced Auditory Serial Addition Test (11)), while in other cases (e.g. imaging) there may not yet be a clear name. Note that the name of the machine is not the name of the outcome measurement instrument; often a machine can be used to measure a variety of parameters (e.g. Greyscale ultrasound [to measure] synovial thickening (synovial hypertrophy) or Doppler ultrasound [to measure] increased blood flow (Synovial hyperemia) (19)), or a pathological entity can be measured by different types of images (for example, enthesitis measured by ultrasound (17) or by MRI (20)). We recommend to include the type of measurement (e.g. ultrasound) in combination with the entity measured as the name of the score (e.g. ultrasound enthesitis score).

Element 2. The version of the outcome measurement instrument or way of operationalization of the measurement protocol

Details on the version, and operationalization of the outcome measurement instrument should be extracted. Details on specific version refer the e.g. the length of the task (e.g. the 2-, 6- or 12-minute walking test (21)), or the number of items included in the version (e.g. Doloplus-1 or Doloplus-2 (13)), or the language used (the English (21) or Dutch version (22) of the 6-minute walk test).

Choices in how the measurement protocol was operationalized may affect the measurement, and should thus be made explicit. Specifically, the components that are potential sources of variation, need to be listed, for example, specific characteristics of the equipment used (e.g. brand and type of the machine), and characteristics of the professionals involved in the measurement (e.g. background and experiences). The taxonomy of the components of measurement instruments (see chapter 2.1) can be used for this.

Element 2 refers to components known or expected to influence the score that are not the object of study. To eliminate the influence of these potential sources of variation on the scores obtained, these components should have been restricted or standardized in the study. For example, if it is expected that different types or brands of machines may interfere with the score, only one type and brand of a machine is used (and reported). In the study by Skeie et al (2015) only the Medison Accuvix V10 ultrasound scanner with a 3–7 MHz curvilinear probe was used (19) – in other words, the brand and type of machine and probe was standardized. Moreover, chiropractors with respectively 4 and 8 years of experienced in diagnostic ultrasound for the musculoskeletal system, and with a
postgraduate diploma in diagnostic ultrasound were involved in the measurements (19). Thus, the background of the raters was restricted to a specific profession (i.e. chiropractors) with specific duration of expertise (4/8 years in diagnostic ultrasound) having received specific training.

In addition, in some cases the instrument procedure requires multiple readings, and a summary statistic (usually the mean, but sometimes the median, maximum or minimum) is calculated as or used to assign the final score (i.e. the results of the measurement). A well-known example is blood pressure measurement in the clinic.¹ How the measurement is taken, should be specified, as it is needed to assess standards 7 (see chapter 3).

For people familiar with the terminology of the Generalizability Theory, the version or the way of operationalization of the measurement instrument refers to the facets of stratification, where patients (i.e. the object of measurement) are nested in a facet (23).

Element 3. The construct measured by the measurement instrument

To identify exactly which outcome measurement instrument was studied, we recommend to extract the construct measured, unless it is clear from the given name. The construct refers to what is being measured, i.e. the ‘aspect of health’. It is also referred to as the ‘concept of interest’ or the ‘intended objective to be measured’. When the measurement instrument does not have a name, identifying the construct can help to fully characterize the outcome measurement instrument (which we also recommend to mention in the name, i.e. element 1). Table 5 provides some examples.

Note that a study on reliability or measurement error does not provide information about whether indeed the construct is being measured, for that you need validity and accuracy studies.

¹ To measure blood pressure, the technician first palpates the radial artery, inflates the cuff until the pulse disappears, inflates an extra 20-30 mm Hg, and then slowly deflates until the pulse reappears. The pressure is noted, and the measurement begins: first, the stethoscope is placed on the brachial artery just medial and above the cubital fold. Then the cuff is reinflated. The pressure is quickly increased to 30 mm Hg above the previous reading, and then slowly deflated until the pulse sounds are detected (systolic blood pressure, measured in 2 mm increments), then further deflated until the sounds disappear (diastolic blood pressure). The cuff is fully deflated, then inflated again to repeat the measurement.
Table 5. Examples of elements 1, 2, and 3.

<table>
<thead>
<tr>
<th>Element 1: name</th>
<th>Element 2: version/ operationalization</th>
<th>Element 3: construct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nine hole peg test (24)</td>
<td>A wooden or plastic board with 9 holes (10 mm diameter, 15 mm depth), placed apart by 32 mm (25)</td>
<td>Finger dexterity</td>
</tr>
<tr>
<td>Ultrasound enthesitis score</td>
<td>Sonography images obtained by experienced sonographers using the Esaote Technos MPX machine</td>
<td>Enthesitis</td>
</tr>
<tr>
<td>HbA1c value based on immune-turbidimetry (12)</td>
<td>Turbidimetric inhibition immunoassay (TINIA), including 2 reagens (i.e. anti-HbA1c antibody (R1), and buffer/polyhapten reagent (R2)); Tetradecyltrimethylammonium bromide (TTAB) is detergent; Roche/Hitachi cobas c systems.</td>
<td>HbA1c (glycated haemoglobin)</td>
</tr>
</tbody>
</table>

Element 4. Specification of the measurement property of interest

When the measurement property of interest is reliability, the study will report relative parameters such as an ICC, Generalizability coefficient \( \varphi \), or Kappa \( \kappa \).

When the measurement property of interest is measurement error, the study will report absolute parameters, either expressed in the unit of measurement, such as SEM, LOA or SDC, or expressed as agreement or misclassification, e.g. the percentage specific agreement.

We recommend to use the COSMIN terminology to determine whether a study assessed reliability or measurement error, regardless of the terms used in the article, because confusion persists about the correct application of these terms. For example, when in a particular article it is stated that ‘reliability’ was assessed, but the standard error of measurement (SEM) or the limits of agreement are reported, the result of that study should be considered as evidence for measurement error (26). When an author states to have evaluated ‘agreement between raters’ using the kappa statistic, the result of this study refers to the reliability of the outcome measurement instrument (27).
Element 5. Specification of the components of the measurement instrument that will be repeated. (Figure 1)

It should be extracted whether the interest of the study is in the reliability or measurement error of the whole measurement procedure (see Figure 1, study A), or only in part of the measurement procedure (see Figure 1, study B). For example, based on an static image that was made once for a patient, only the assignment of the score was repeated, or the performance of a task of each patient was videotaped, and only the last component (i.e. assignment of the scores) is repeated.

![Figure 1](image_url)

**Figure 1.** Which part of the measurement is repeated.

Element 6. Specification of the components of the measurement instrument that will be varied

The component of the measurement instrument that is being varied across the measurements is the main focus of the study. Examples are time or occasion (test-retest, or intra-rater), the professionals (inter-rater), or the machines (inter-machine or inter-device) (28). For example, in Figure 1 raters are varied: rater A conducts the first measurement and rater B conducts the second measurement for each patients.
In the design of the study one or more sources can be considered. For example, both the machine and the rater who conducts the whole measurement are varied across the repeated measurements (see Figure 2, study A). The taxonomies of components of measurement instruments (see chapter 2.1) can be used to consider various potential sources of variation.

Figure 2. Designs in which components are varied across repeated measurements

Alternatively, the researchers can assume that a component (e.g. preparation or assignment of the score) is 'stable', in other words, that the rater who prepares the measurement or who assigns the score will not introduce error in this part of the measurement (indicated in grey in Figure 2 study B and C), and investigate only the influence of the components (e.g.) equipment, preparation, collection of raw data and data processing and storage.

In the designs shown in Figure 1 and 2 we assume that all patients were measured this way. This is called a crossed design (29). However, so-called nested designs are possible, too (see Figure 3). In these designs, part of the patients are measured following measurement conditions A and other patients are measured using measurement conditions B. In Figure 3 a nested inter-rater reliability design is shown, where some of the patients are measured first by rater A and next by rater B (i.e. measurement condition A), while other patients are measured first by Rater C and next by rater D (i.e. measurement condition B), etc. These designs are appropriate to use, and in the calculation of the ICC, this could be taken into account. For example, by calculating
variance components per measurement condition, and next pool these variance components (weighted by sample size) across the measurement conditions (e.g. (30)), or by using a one-way random effects model (31).

Figure 3. Nested inter-rater reliability design.

For people familiar with the terminology of the Generalizability Theory, the components that are being varied across measurements are called the *random or fixed facets of Generalizability* (23).

Element 7. Patient population

The reliability depends on the homogeneity or heterogeneity of the study population. Therefore, the sample (and its subgroups) included in the study should be extracted and assessed by the user of this tool. In the study by Skeie et al (2015) the recruited sample consisted of low back patients, patients with other spinal complaints, but also of pain-free subjects. This latter group could have increased the variance between patients, and subsequently, influenced the results (i.e. increased the ICC) of the reliability study.

In the COSMIN methodology we use the word patient. However, sometimes the study population of interest consists of healthy individuals, body structures (e.g. joints, kidneys), clinicians or caregivers. In these cases, the word patient should be read as e.g. healthy person or caregiver.

For people familiar with the terminology of the Generalizability Theory, the patient population refers to the *object of measurement or the facets of differentiation* (23).
2.3 Example of how to use Part A of the COSMIN Risk of Bias tool to assess the quality of a study by Skeie et al. (2015)

In this chapter we provide an example of how to use the COSMIN tool – Part A using a paper by Skeie et al. (19). To get a full understanding of the study, we recommend to first read the introduction and method section of the paper. In this paper four different studies are described. Here we use the first two sub studies, and provide a summary of these two studies.

In this paper, the lumbar multifidus muscle (LMM) thickness score (study 1) and contraction score (study 2) was investigated by ultrasound. The measurement proceeds as follows: a patient is asked to lay down in a specific position, and the probe is placed on a very specific body part. This yields an on-screen image. Subsequently, a marker is placed on a specific structure (i.e. the apex of the facet joint) identified on the image. In study 1, a still image is recorded, and the first rater places the second marker on another specific structure (i.e. processus mammillaris) on this image, and measures the distance between the markers with the calliper software. The two markers correspond with the thickness of the LMM. The first rater repeats the second marker placement and distance measurement on the still image twice, for a total of three measurements. The patient leaves. Next, based on the very same still image (with only the first marker visible) a second rater places the second marker on the screen and measures the distance a total of three times. Next, all data is transferred to a separate paper by rater 1 who calculates a mean value per patient per rater. This mean value is the LMM thickness score. The repeated placement of the second marker on the still image and application of the caliper tool to measure the distance between the two markers is part of one measurement (19). This procedure is depicted in Figure 3, study 1.

Figure 3. Study designs of Skeie et al.
In study 2, for each patient each of the raters independently generated one image of the LMM in the resting state and one image of the LMM in contracted state. Using a split-screen of the two still images of both states, each rater measured thickness (i.e. caliper-assessed distance between the markers) of the two states three times. Next, rater 1 transferred the data to a separate paper and calculated mean values of the thickness of each state. Next, rater 1 calculated the ‘LMM contraction score’ as the exact change in thickness (contracted LMM minus resting LMM) (19). This procedure is depicted in Figure 3, study 2.

Based on the thorough elaboration of the study performed and described by Skeie and colleagues, we extract the elements of a comprehensive research question.

Table 6. Example of how to use Part A of the COSMIN Risk of Bias tool based on the study by Skeie (19).

<table>
<thead>
<tr>
<th>Element</th>
<th>Instruction</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Name of the instrument</td>
<td>Alternatively: type of instrument and parameter</td>
<td>Ultrasound measurement of the lumbar multifidus muscle (LMM) thickness score</td>
<td>Ultrasound measurement of the LMM contraction score</td>
</tr>
<tr>
<td>2. Version or way of operationalization</td>
<td>All relevant components that are known or expected to influence the score, and which are standardized or restricted (facet of stratification (23))</td>
<td>Equipment: Medison Accuvix V10 ultrasound scanner with a 3–7 MHz curvilinear probe; Preparatory actions: two chiropractors with 4 respectively 8 years of experience in diagnostic ultrasound for the musculoskeletal system, with a postgraduate diploma in diagnostic ultrasound; still on-screen images were obtained with the subjects in a prone position with a pillow placed under the abdomen to flatten the lumbar lordosis. Preparation: Image was on-screen generated and a marker was placed on the image on the mamillary process of the level to be measured. Unprocessed data collection: The second marker was placed on the on-screen image, and the distance was computed by the calliper software. This part was repeated three times.</td>
<td>Preparation: In resting position, an image was on-screen generated and a marker was placed on the image on the mamillary process of the level to be measured. Next, in contracted state (LMM contraction was induced by a contralateral arm lifting task), an image was on-screen generated, too, and a marker was placed on the image.</td>
</tr>
<tr>
<td>Element</td>
<td>Instruction</td>
<td>Study 1</td>
<td>Study 2</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Data processing and storage:</td>
<td>Data is transferred to a separate paper by rater 1.</td>
<td><strong>Unprocessed data collection:</strong></td>
<td>based on the split-screen of both images, the second marker was placed on each image, and the distance (per image) was calculated by the calliper software. This part was repeated three times.</td>
</tr>
<tr>
<td>Assignment of the score:</td>
<td>Rater 1 calculated a mean value per patient per rater.</td>
<td><strong>Assignment of the score:</strong></td>
<td>Rater 1 calculates a mean value per patient per rater for both states. Next, the rater calculated the ‘LMM contraction score’ as the exact change in thickness (contracted LMM minus resting LMM).</td>
</tr>
<tr>
<td>3. Construct</td>
<td>Description of what is being measured</td>
<td>LMM thickness</td>
<td>LMM contraction, which is change in LMM thickness in contracted and resting state (contracted LMM minus resting LMM).</td>
</tr>
<tr>
<td>4. Measurement property</td>
<td>Reliability and measurement error</td>
<td>Reliability and measurement error</td>
<td></td>
</tr>
<tr>
<td>5. Components that will be repeated</td>
<td>Either the whole measurement (i.e. all components) or the assignment of the score (i.e. last component)</td>
<td>The whole measurement will be repeated. However, the focus of interest in on the unprocessed data collection: placing of the second marker on the on-screen image (mean of three times).</td>
<td>The whole measurement will be repeated. However, the focus of interest in on the preparation (i.e. preparation and generation of images in the resting and contracted states, and the placing of the first marker), and on the unprocessed data collection (placing of the</td>
</tr>
<tr>
<td>Element</td>
<td>Instruction</td>
<td>Study 1</td>
<td>Study 2</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>6. Source(s) of variation varied</td>
<td>Components which is varied across the measurements (i.e. focus of analysis; facet of generalizability (23))</td>
<td>Raters (n=2; inter-rater reliability)</td>
<td>Raters (n=2; inter-rater reliability)</td>
</tr>
<tr>
<td>7. Patient population</td>
<td>(i.e. facet of differentiation (23))</td>
<td>LBP patients, patients with other spinal complaints such as mid back pain, neck pain, and/or extremity pain, and pain-free subjects (n=30 in each experiment, total n=120)</td>
<td></td>
</tr>
</tbody>
</table>

Based on the extracted information, a comprehensive research question can be formulated as:

**Study 1**: What is the inter-rater reliability of the data collection phase of the lumbar multifidus muscle (LMM) thickness score based on the mean of three marked distance with the calliper software on a still image of the ultrasound measurement, measured using the Medison Accuvix V10 ultrasound scanner with a 3–7 MHz curvilinear probe by post-graduate experienced chiropractors, in LBP patients, patients with other spinal complaints such as mid back pain, neck pain, and/or extremity pain, and pain-free subjects?

**Study 2**: What is the inter-rater reliability of preparing, generating, and data collection phases of the lumbar multifidus muscle (LMM) contraction score, based on the mean of three marked distance with the calliper software on an on-screen image in resting and in contraction state of the ultrasound measurement, measured using the Medison Accuvix V10 ultrasound scanner with a 3–7 MHz curvilinear probe by post-graduate experienced chiropractors, in LBP patients, patients with other spinal complaints such as mid back pain, neck pain, and/or extremity pain, and pain-free subjects?

Please, note that we do not recommend to report the research question always as this in one long question. Though, we consider it very useful to describe all this information clearly, e.g. in the method section of a paper.
3. Part B. Assessing the risk of bias of a study on reliability or measurement error

Part B of the COSMIN Risk of Bias tool contains two boxes with standards that can be used to determine whether the result of a study on reliability or measurement error, respectively, can be trusted. Standards refer to the design requirements of the study or to the preferred statistical methods.

The standards 1 to 5 in both boxes refer to design requirements. These standards are the same for studies on reliability and for studies on measurement error, as the same design can be used for assessing both measurement properties.

Three standards refer to the preferred statistical methods for studies on reliability and two standards refer to the preferred statistical methods for studies on measurement error. In the COSMIN Risk of Bias tool, we included standards concerning the preferred statistical methods that are appropriate to use when evaluating reliability or measurement error of outcome measurement instruments (see also section 1.6). Other methods may be appropriate to use as well (for example bi-factor models or Multi-Trait Multi-Method (MTMM) analyses, or newly developed methods). It is not our intention to comprehensively describe all possible statistical methods, rather to describe the adequate methods that are commonly used in the literature.

Each box also contains a standard asking if there were any other important methodological flaws that are not covered by the other standards (standard 6), but that may have led to biased results or conclusions. Some flaws are rather uncommon, and therefore, do not justify a separate standard. In chapter 3.1 we provide several examples for these flaws.

Each standard will be scored on a four-point rating system (i.e. ‘very good’, ‘adequate’, ‘doubtful’, or ‘inadequate’) in line with the COSMIN Risk of Bias checklist for Patient-Reported Outcome Measures (PROMs) (1). Subsequently, the lowest rating given in a box determines the final rating, i.e. the quality of the study (this is called the worst-score-counts method (18) to determine the risk of bias). Sometimes a response option is indicated in grey, meaning that the response option is not applicable for the standard, and users should choose between the other options. Final, some standards can be rated as ‘not applicable’.

In general, a standard on a design requirement is rated as ‘very good’ when there is evidence or convincing arguments were provided that the standard is met; ‘adequate’ when it is assumable, although not explicitly described, that the standard is met; ‘doubtful’ when it is unclear that the standard is met; and ‘inadequate’ when there is evidence that the standard is not met (18).

A standard about preferred statistical methods is in general rated as ‘very good’ when a preferred method was optimally used; ‘adequate’ when the preferred method was used,
but it was not optimally used, ‘\textit{doubtful}’ when it is unclear if a preferred method was used, and ‘\textit{inadequate}’ when the statistical methods used are considered inadequate.

The boxes for reliability and measurement error, respectively, can be found here. Below, an elaboration of each standard is described for reliability (chapter 3.1) and measurement error (chapter 3.2). In chapter 3.3 we provide an example for rating the box on reliability in the study by Skeie, that was also used as an example in chapter 2.3.
3.1 Elaboration on standards for studies on reliability

The box on reliability contains five standards about design requirements, one standard ‘other flaws’ and three standards about preferred statistical methods. For each standard we give suggestions for how to rate the standard.

Standard 1. Stability of the patient

<table>
<thead>
<tr>
<th>Were patients stable in the time between the repeated measurements on the construct to be measured?</th>
<th>very good (evidence provided)</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate (evidence provided)</th>
<th>NA applicable</th>
</tr>
</thead>
</table>

**Elaboration:** Patients should be stable with regard to the construct to be measured between the repeated measurements. When an intervention such as surgery or medication is given in the interim period, it is likely that (many of) the patients have changed on the construct to be measured. In other words, they are not stable – and the standard should be rated as ‘inadequate’. When the aim is to assess the reliability of the assignment of the score, e.g. using static images or videos of the performance of a task as object of interest (see Figure 1 study 2 – page 24), this standard is not applicable as the images and videos were acquired only once. Furthermore, the measurement can interfere with the stability of the patient. For example, there should be enough time for patients to recover from experienced pain or fatigue between repeated measurements and permit patients to return to their initial state. If not, the standard should be rated as ‘doubtful’, as it is unclear whether the patients are stable on the construct to be measured. When evidence or convincing arguments are provided that the patients were stable, the standard is scored ‘very good’.

Standard 2: Time interval

<table>
<thead>
<tr>
<th>Was the time interval between the measurements appropriate?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
</table>

**Elaboration:** The time interval between the measurements must be appropriate. The definition of “appropriate” depends on the construct to be measured and the study population. The time interval should be long enough to prevent recall bias of previous scores in case of intra-rater reliability, and short enough to ensure that patients have not changed on the construct to be measured. For example synovitis can change in a few days, while a change in cartilage or bone status would take a few months.
Standard 3. Similar measurement conditions

<table>
<thead>
<tr>
<th>Were the measurement conditions similar for the measurements – except for the condition being evaluated as a source of variation?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (evidence provided)</td>
<td>Reasons to assume standard was met, OR change was unavoidable</td>
<td>Unclear</td>
<td>No (evidence provided)</td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** Each repeated measurement should be conducted with the same measurement protocol – except for the source of variation that was intentionally varied, i.e. element 6 of the comprehensive research question (see chapter 2.2). For example, if the aim was to understand the variation due to different raters (i.e. inter-rater reliability), only the raters should be varied. Other concomitant sources of variation (i.e. element 2 of the comprehensive research question, see chapter 2.2) should be kept similar. Was the study up to standard? Were all equipment, preparatory actions, the environmental conditions (e.g. temperature), and methods of processing the same in both measurements? For example, when the patients are very likely to show a learning effect (for example on a performance-based test), the absence of a familiarization session should yield a rating of doubtful or inadequate on this standard, as the first measurement can then be considered to be the familiarization session, and the measurement conditions are not the same. A description of similarity of the measurement conditions of the repeated measurements can be considered as evidence.

Standards 4. Administration of measurements

In instruments that do not involve biological sampling, the administration refers to the components 'Collection of raw data' and 'Data processing and storage' (see chapter 2.1). In instruments involving biological sampling, it refers to the components 'Collection of biological sampling' and 'Biological sampling processing and storage' (see chapter 2.1).

<table>
<thead>
<tr>
<th>Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (evidence provided)</td>
<td>Reasons to assume standard was met</td>
<td>Unclear</td>
<td>No (evidence provided)</td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** All measurements should be administered by the professional(s) involved without them having knowledge of the scores or values of other repeated measurements on the same outcome measurement instrument. This means that the measurements should all be administered without knowledge of the prior (e.g. in case of an intra-rater reliability study) or other (e.g. in case of an inter-rater reliability study) score(s) or value(s) on the instrument of interest.
The rating of this standard is rather subjective. For example, if in a study the raters independently administered the measurement, and none were involved in the care of the patients (making it very unlikely that the raters received additional information of the patients including knowledge of the score(s) of other repeated measurements), this can be considered as ‘evidence provided’, and the rating is ‘very good’. When the other score is known to the professional while administering the repeated measurement, it may influence the way the measurement is administered. For example, with a severe score obtained with an imaging technique, the repeated measurement can be administered more carefully, and more time can be used to look at the patient. If it is known that this was the case, the rating is ‘inadequate’. When there is no explicit description, but it seems very unlikely that the raters knew the scores or values of other repeated measurements, it can be rated as ‘adequate’, or ‘doubtful’. In some situations this standard is not applicable, i.e. when the administration (i.e. collection of the raw material or biological sample, data or sampling processing and storage) is not repeated in the study, but only the assignment of the score or the determination of the value (see for examples chapter 2.2 element 5 of the comprehensive research question, or Figure 1 study 2).

**Standard 5. Assignment of the score or determination of the biological value**

<table>
<thead>
<tr>
<th>Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes (evidence provided)</td>
<td>Reasons to assume standard was met</td>
<td>Unclear</td>
<td>No (evidence provided)</td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** The scores on all measurements should be assigned or values should be determined by the professional(s) involved without them having knowledge of the scores or values of other repeated measurements. This means that assigning a score to a measurement or determining the value of a biological sample should be done without knowledge of the prior (e.g. in case of an intra-rater reliability study) or other (e.g. in case of an inter-rater reliability study) score(s) or value(s) on the instrument of interest. Although part of the determination of the value of a biological sample can be an automatic step, there may be human action required to do this determination. For example, an urine pH level test to measure the acidity or alkalinity of urine where the color of the strip is interpreted by the professional. The rating is similarly as explained for standard 4.
Standard 6. Other important flaws

<table>
<thead>
<tr>
<th></th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Were there any other important flaws in the design or statistical methods of the study?</td>
<td>No</td>
<td>Minor other methodological flaws</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** This standard is included because there might be uncommon design flaws that are not covered by other standards but that may cause additional risk of bias. Below, some examples are provided.

When various professionals are involved in the measurement instrument, and one of the professionals is the attending physician of the patient, this physician has (much) more information about the patient than the other professionals. In some situations – depending on the aim of the study and the specific construct to be measured – this could be considered a flaw because of the influence on the scores obtained.

In the previous chapter we saw in the example of Skeie that part of the sample comprised healthy patients, whereas the authors were ultimately interested in these measurements in low back pain patients (19). As this will increase the variance between patients, and it will increase the results of the study (i.e. the ICC or G Coefficient). Depending on where this study sits in the development of the instrument, this could be deemed proper (when the full range of the scores is not yet known) or an important flaw when the purpose is to determine the reliability of measurement in the clinical setting of low back pain.

A final example refers to the use of the ICC model for average scores. Although discussed under standard 7 for reliability, it may be that the ICC for the mean score of the measurements is reported, whereas in clinical practice the single score is used. Depending on the purpose of the study this can be proper (when the mean score is going to be used in future research) or an important flaw when the study is aimed at proving reliability on clinical practice (where the single score is used).

It is up to the user of the COSMIN Risk of Bias tool whether a flaw is considered minor (and is rated as ‘doubtful’) or important (and is rated as ‘inadequate’). The scores of the other flaws are included in the overall score/rating based on the worst score counts principle.
Standard 7: Preferred statistical methods for continuous scores

<table>
<thead>
<tr>
<th>For continuous scores: was an intraclass correlation coefficient (ICC) calculated?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>Inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC calculated; the model or formula was described, and matches the study design and the data</td>
<td>ICC calculated but model or formula was not described or does not optimally match the study design OR Pearson or Spearman correlation coefficient calculated WITH evidence provided that no systematic difference between measurements has occurred</td>
<td>Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic difference between measurements has occurred OR WITH evidence provided that systematic difference between measurements has occurred</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Elaboration: For continuous scores the intraclass correlation coefficient (ICC) is preferred to evaluate reliability. ICCs are a family of statistical parameters, including Generalizability (G) coefficients, and Decision (D) coefficients.

To get a “very good” rating, the ICC model used in the reliability study should match the study design (and the aim) of the study that is being assessed. Therefore, the model or formula of the ICC or G Coefficient used should be described. It should be clear, e.g. whether a crossed or nested design was used (see also page 25/26), or whether a one-way random effects model, two- or three-way random or mixed effects model was used. Next, it should be compared to the study design using the extracted information from Part A, and determined whether the ICC or G Coefficient used indeed matches the study design.

The ICC based on the two-way mixed effects model of consistency (31) (also referred to as ICC model 3.1 (32)), and the Pearson or Spearman correlation coefficient do not take a systematic difference between the repeated measurements into account, and are therefore considered less appropriate, as it can lead to overestimating the reliability. Therefore, based on information of a systematic difference between the source of variation considered (e.g. raters) either ‘adequate’ (when no or very little systematic difference occurred), or ‘doubtful’ (when there was a systematic difference between e.g. the raters) can be rated.

When the study was designed to investigate a specific source of variation (e.g. inter-rater), and the systematic differences between this source of variation in the repeated measurements was taken into account in the formula (for example, by using the ICC random effects model for agreement (31), also referred to as Model 2.1 (32) or the \( \varphi \) coefficient (see e.g. (23)), the study can be rated as ‘very good’.

When a study is designed without any specific source of variation is considered, the appropriate ICC model is a one-way random effects model (31). In this situation the use
of a one-way random effects model can be rated as 'very good', while the use of other models can be rated as 'adequate'.

Next, the ICC can be calculated for a single measurement or an average measurement (31). If a single measurement is normally used in clinical practice or trials (and not the average score of multiple measurements, such is done by a blood pressure measurement), the ICC for single measures should have been calculated. The ICC average refers to the reliability of the averaged score of the measurements, and refers to the use of the averaged score on repeated measurements. When the ICC for average measures is reported, in the situation that usually a single measurement is taken, we recommend this standard to be rated as ‘adequate’, as the model does not optimally match the design of the study. However, we also recommend in this situation, to rate standard 6 (i.e. other flaws), as ‘doubtful’ or even ‘inadequate’ (see also the example at standard 6).

Moreover, to get a ‘very good’ rating, the described ICC or G coefficient model or formula should match the data. If there is a (known) problem with normal distribution of the data (normality) which is not properly taken into account, the study could be rated as ‘adequate’ instead of ‘very good’.

It is impossible to describe all other flaws here, Therefore it is up to the user of the COSMIN Risk of Bias tool to decide how the identified flaw should be scored. Relevant question in this regard is how certain and how large the influence is on the study result.

### Standard 8: Preferred statistical methods for ordinal scores

<table>
<thead>
<tr>
<th>For ordinal scores: was a (weighted) kappa calculated?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa calculated; the weighting scheme was described, and matches the study design and the data</td>
<td>Kappa calculated, but weighting scheme not described or does not optimally match the study design</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** To assess reliability for ordinal scores, Cohen’s kappa (33-35) is considered the preferred statistical parameter. No better alternative is known (4, 36).

Information on the specific kappa used should be described in terms of whether a weighting scheme was used and which scheme was used. Unweighted kappa considers any misclassification equally inappropriate. However, a misclassification of two adjacent categories may be less erroneous as a misclassification of categories that are more apart from each other. A weighted kappa takes this into account (e.g. using linear or quadratic weights (37)). If the goal of the study was to consider any misclassification as equally important, and it was stated that the unweighted kappa was used, this standard can be rated a ‘very good’. However, in other situation (e.g. misclassification of categories more
apart from each other is a bigger problem that misclassification of adjacent categories) a specific weighting scheme is more preferred. If unweighted kappa calculated in that case the standard could be rated as ‘adequate’.

**Standard 9: Preferred statistical methods for dichotomous or nominal scores**

<table>
<thead>
<tr>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?</td>
<td>Kappa calculated for each category against the other categories combined</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** A study on reliability of an outcome measurement instrument with dichotomous or nominal scores gets a ‘very good’ score, when an unweighted kappa was calculated of each category against the other categories (33).
3.2 Elaboration on standards for studies on measurement error

Standards 1 to 6 of the box for standards for studies on measurement error are the same as for studies on reliability. For an elaboration on each of the standards, please see above.

Standard 7: Preferred statistical methods for continuous scores

<table>
<thead>
<tr>
<th>For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEM, SDC, LoA or CV calculated; the model or formula for the SEM/SDC is described; it matches the reviewer constructed research question and the data</td>
<td>SEM, SDC, LoA or CV calculated, but the model or formula is not described or does not optimally match the reviewer constructed research question and evidence provided that no systematic difference has occurred</td>
<td>SEM_{consistency}</td>
<td>SEM, SDC_{consistency} or LoA or CV calculated, but without knowledge about systematic difference or with evidence provided that systematic difference has occurred</td>
<td>SEM calculated based on Cronbach’s alpha, OR using SD from another population</td>
</tr>
</tbody>
</table>

Elaboration: For continuous scores preferred measures for the measurement error of a single score are the SEM, LoA or the Coefficient of Variation (CV); the SDC is preferred as a measure for change scores.

Different formulas can be used to come to calculate these various measures. Therefore, we will first describe their formulas. Subsequently, we will explain the standard for studies using SEM and SDC derived from variance components analyses. Next, we will discuss LoA, SEM and SDC using the SD_{difference}. We will explain when ignoring the influence of the source of variation is appropriate. And last, we will discuss some other methods used, including the CV.

Measures that take all error into account, including the systematic difference between repeated measurements, based on a one-way or two-way effects model, are:

\[ SEM_{one-way\ random\ effects\ model} = \sqrt{\sigma_{error}} \]  \hspace{1cm} (1)

\[ SEM_{agreement} = \sqrt{\sigma_{rater} + \sigma_{random}} \]  \hspace{1cm} (2)

\[ SDC_{agreement} = 1.96 * \sqrt{2} * SEM_{agreement} = 1.96 * \sqrt{2} * \sqrt{\sigma_{rater} + \sigma_{random}} \]  \hspace{1cm} (3)
Measures that do not take the systematic difference between repeated measurements into account:

\[ SEM_{\text{consistency}} = \sqrt{\sigma_{\text{random}}} \]  
\[ SDC_{\text{consistency}} = 1.96 \times \sqrt{2} \times SEM_{\text{consistency}} = 1.96 \times \sqrt{2} \times \sqrt{\sigma_{\text{random}}} \]  
\[ SEM_{\text{consistency}} = \frac{SD_{\text{diff}}}{\sqrt{2}} \]  
\[ SDC_{\text{consistency}} = 1.96 \times \sqrt{2} \times SEM_{\text{consistency}} = 1.96 \times \sqrt{2} \times \frac{SD_{\text{diff}}}{\sqrt{2}} \]  

Limits of Agreement = \( d \pm 1.96 \times SD_{\text{difference}} \)  
\[ SDC_{\text{consistency}} = 1.96 \times SD_{\text{difference}} \]

To get a 'very good' rating, the formula used should match the study design (and the aim) of the study that is being assessed. Therefore, it should be clear what the aim is, and which measure or which formula was used in the study being assessed.

Measurement error derived from variance components analyses (formulas 1-5)
The specific model used should be clearly described, e.g. whether a one-way random effects model, or a two- or three-way random or mixed effects model was used, and whether all error (except from the variance due to variation between patients) was included in the calculation of the measurement error, or whether the systematic error between the source of variation that is being varied in the design is ignored (i.e. as occurred when calculating \( SEM_{\text{consistency}} \) for single scores (formula 4) and \( SDC_{\text{consistency}} \) for change scores (formula 5)). Next, it should be compared to the study design using the extracted information about the comprehensive research question (see Part A of the tool), and determined whether the method used indeed matches the study design. In other words, when the aim of the study was to assess the measurement error of a single score of any measurement taken in clinical practice of trials, the aim is to generalize the results beyond (e.g.) the specific raters involved in the study. In this case, the systematic error between raters should be taken into account; the raters (in this example) should be considered random; and all error should be taken into account (i.e. formulas 1-3) to match the design of the study (and this is rated 'very good').

If in this case, (with the aim to generalize beyond the specific raters) the \( SEM_{\text{consistency}} \) (formula 4) or \( SDC_{\text{consistency}} \) (formula 5) was calculated (i.e. ignoring a systematic
difference between raters), evidence should be provided that no (or only very small) systematic difference has occurred between the raters. In case of no or very small differences the standard can be rated as ‘adequate’, as the SEMagreement (formula 2) and SEMconsistency (formula 4), or SDCagreement (formula 3) and SDCconsistency (formula 5) will be the same or very close. If it is unclear whether systematic differences occurred (because it was not reported), the standard is rated as ‘doubtful’.

Measurement error derived from the SDdifference (formulas 6-9)

The measurement error of a single score or a change score can also be calculated using the SDdifference. This refers to the standard deviation of the difference of the scores on the repeated measurements (38, 39). In a Bland and Altman plot two repeated measurements per patient are plotted: on the x-axes the mean score of the two measurements, and on the y-axes the mean difference between the repeated measurements (39). Although the plot is designed in such a way that systematic differences can easily be seen (i.e. the line of the mean differences in scores, and the asymmetrically located limits of agreement around the zero), the systematic difference is disregarded when the SDC is calculated from these limits (resulting in the SDCconsistency). Therefore, if a (large) systematic error between the repeated measurements occurred, while the aim of the study is to generalize beyond the specific source of variation (e.g. raters), the standard should be rated as ‘doubtful’, as the results of the study is underestimating the measurement error.

When is a measure of consistency (formulas 4-9) appropriate?

Sometime, the source of variation that is being varied across the measurements is considered to be fixed in a study. This means that the aim of the study is not to generalize beyond the specific study objects included in the study. For example, in a study only two raters are considered (e.g. the raters Myrthe and Brechtje), and the aim of the study is whether these two raters will come to equal scores (e.g. because they will be the only two raters involved in the measurements for a specific trial). If a systematic error occurs between Myrthe and Brechtje (e.g. Myrthe systematically scores 5 points higher compared to Brechtje), the scores obtained in the trial can easily be adjusted by extracting 5 points of each measurement obtained by Myrthe. In this study, the source of variation ‘rater’ is deemed irrelevant (31), as the systematic difference will be adjusted later on when using the instrument by either Myrthe or Brechtje. In this specific situation, the SEMconsistency, SDCconsistency or the limits of agreement match the aim and design of the study, so it can be rated as ‘very good’. However, these results cannot be generalized to other raters, as ‘rater’ was considered fixed. Therefore, the study is less relevant in other situations, especially when there is a systematic difference between the raters.
Measurement error calculated using the formula \( SD \times \sqrt{1 - ICC} \)

There is another formula which is sometimes used to calculate the SEM from the ICC formula: \( SEM = SD \times \sqrt{1 - ICC} \) (40). The standard deviation refers to the SD\(_{\text{pooled}}\) of the sample, that is of SD\(_{\text{test}}\) and SD\(_{\text{retest}}\). Using this formula is only justified if the data for ICC and SD are derived from the same study. When the SD is based on another population, this is considered inadequate, as the SD of this other population may be smaller, and subsequently, the measurement error is smaller.

Moreover, sometimes the Cronbach’s alpha is inserted in the formula instead of the ICC. This is considered inadequate, as this measure is based on one full-scale measurement where items are considered as the repeated measurements, instead of at least two full-scale measurements using the total score in the calculation of the SEM. Often Cronbach’s alpha is higher than ICC’s based on repeated measurements, thus leading to smaller SEM values. By rating this inadequate, the result of this study can still be considered, however, it is considered to be less trustworthy. Moreover, Cronbach’s alpha is sometimes used inadequately, because it is calculated for a scale that is not unidimensional, or based on a formative model. In such cases the Cronbach’s alpha cannot be interpreted.

Other parameters that are based on single measurements, such as the person separation index (or other IRT-based measurement error measures) or the Omega, are not covered by the measurement error according to the COSMIN taxonomy, but by internal consistency.

The Coefficient of variation

Coefficient of variation (CV) is also a parameter of measurement error. It is often used in physics and to present the measurement error of a device. When developing a new device the measurement error is assessed by measuring a fixed sample many (e.g. 50) times. The SD of these measurements is the standard error of measurements. Often the measurement error increases with higher values. For these situation CV is a suitable measure, as CV expresses the SD as percentage of the mean value: in formula \( CV = \frac{SD}{\text{mean}} \). Usually, it is expressed in percentage, for example, the measurement error is 2% of the measured value.

The assumption underlying CV is that the CV gives a constant value over all values of the mean, so that the SD is e.g. 2% of the mean value, regardless of a mean value of 10 or 100 or 1000.

In a Bland and Altman plot, we had a contrary assumption, i.e. that the SD of the difference is constant over the mean values, on the X-axis. If the differences are lower with small values and higher with large value the horizontal lines of the limits of agreement give a wrong value: too large for the small values and too small for the large mean values. In that case one should transform the data. Often a natural logarithm or 10 log logarithm transformation is used. This has the advantage that the limits of agreement can be directly expressed in a coefficients of variation (41).
Standard 8: Preferred statistical methods for dichotomous, nominal, or ordinal scores

<table>
<thead>
<tr>
<th>For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?</th>
<th>very good</th>
<th>adequate</th>
<th>doubtful</th>
<th>inadequate</th>
</tr>
</thead>
<tbody>
<tr>
<td>% specific agreement calculated</td>
<td>% agreement calculated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Elaboration:** Often kappa is considered as a measure of agreement, however, kappa is a measure of reliability (42). An appropriate parameter of measurement error (also called agreement) of dichotomous/nominal/ordinal scores is the proportion of specific agreement (42-44). It is a measure that expresses the agreement separately for each category of the score – that is positive and negative ratings agreement in case the score is dichotomous.
3.3 Example of how to use Part B of the COSMIN Risk of Bias tool to assess the quality of a study by Skeie et al. (2015)

In this chapter we provide an example of how to use the COSMIN tool – Part B using again the paper by Skeie et al., (19). To fully understand the explanation in Table 7, we recommend to first read the introduction and method section of the paper, and the summary provided at page 27/28. In this paper four different studies are described. Here we use the first two sub studies.

Table 7. Example of how to use Part B of the COSMIN Risk of Bias tool based on the study by Skeie (19).

<table>
<thead>
<tr>
<th>Standards on design requirements for Reliability and Measurement error</th>
<th>Rating study 1</th>
<th>Rating study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design requirements</strong></td>
<td><strong>Rating study 1</strong></td>
<td><strong>Rating study 2</strong></td>
</tr>
<tr>
<td>1 Were patients stable in the time between the repeated measurements on the construct to be measured?</td>
<td>NA (measurements were based on a still image)</td>
<td>Very good. Measurements were conducted in succession.</td>
</tr>
<tr>
<td>2 Was the time interval between the repeated measurements appropriate?</td>
<td>NA</td>
<td>Very good. The time interval (i.e. the second rater started immediately after the first had completed the procedure) has probably not influenced the scores.</td>
</tr>
<tr>
<td>3 Were the measurement condition similar for the repeated measurements – except for the condition being evaluated as a source of variation?</td>
<td>Very good</td>
<td>Very good</td>
</tr>
<tr>
<td>4 Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?</td>
<td>Very good. None of the previous scores were available</td>
<td>Very good. None of the previous scores were available</td>
</tr>
<tr>
<td>5 Did the professional(s) assign the scores or determined the values without knowledge of the scores or values of other repeated measurement(s) in the same patients?</td>
<td>Very good. None of the previous scores were available</td>
<td>Very good. None of the previous scores were available</td>
</tr>
<tr>
<td>6 Were there any other important flaws in the design or statistical methods of the study?</td>
<td>For reliability: Doubtful. 5 of 30 persons (see Table 1 of the paper) were pain-free subjects, which could have majorly increased the variation between the patients, and subsequently the ICC. For measurement error: very good. Heterogeneity of the sample is considered less a problem, as the variation between patients is not included in the parameter.</td>
<td>For reliability: Very good. (in this study no pain-free persons were included, see Table 1 of the paper)</td>
</tr>
</tbody>
</table>
## Standards on preferred statistical methods for Reliability

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Rating study 1</th>
<th>Rating study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>For continuous scores: was an Intraclass Correlation Coefficient (ICC) calculated?</td>
<td>Adequate. ICC two-way mixed single measures (3.1) and two-way mixed average measures (3.2) were calculated. This is the ICC\textsubscript{consistency}, which does not take the systematic error between raters into account. The study aims to generalize beyond the raters involved, therefore, the raters should not be considered fixed, and the ICC model does not match optimally the research aim and design. Based on the mean of the measurements provided in Table 2, we can conclude that no systematic difference between the raters occurred. The ICC two-way mixed average measures (3.2) refers to the practice in which two raters would measure each patient (with triple placement of second marker), and both final scores were averaged. As this will not be common practice, we will ignore this ICC. The repetition of part of the measurement is already part of one measurement. Based on the mean of the measurements provided in Table 2, we can conclude that no systematic difference between the raters occurred. The ICC two-way mixed average measures (3.2) refers to the practice in which two raters would measure each patient (with triple placement of second marker), and both final scores were averaged. As this will not be common practice, we will ignore this ICC. The repetition of part of the measurement is already part of one measurement.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>For ordinal scores: was a (weighted) Kappa calculated?</td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
<tr>
<td>9</td>
<td>For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?</td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

### Final Risk of Bias rating Reliability studies

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Doubtful</td>
</tr>
</tbody>
</table>

## Standards on preferred statistical methods for Measurement error

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Rating study 1</th>
<th>Rating study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?</td>
<td>Adequate, as the limits of agreement were calculated, while the aim was to generalize beyond the raters included in this study, and probably there was no systematic difference between the raters.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?</td>
<td>Not applicable</td>
<td>Not applicable</td>
</tr>
</tbody>
</table>

### Final Risk of Bias rating study on Measurement error

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adequate</td>
</tr>
</tbody>
</table>
4. Using the COSMIN Risk of Bias tool in a systematic review of outcome measurement instruments

Researchers and clinicians who are deciding on the most suitable outcome measurement instrument for use in their study, can often choose from multiple different instruments. The selection should be based on the evidence of the quality of the outcome measurement instruments (i.e. reliability, validity, and responsiveness), as well as on aspects of feasibility and interpretability. A high-quality systematic review on outcome measurement instruments gives a clear overview of all important aspects to make your choice.

Understanding the quality of the studies and the quality of the measurement instrument under study is a challenging task, specifically for researchers and clinicians who are less familiar with the methodology to evaluate all measurement properties. Therefore, in 2018, we (COSMIN initiative) published a thorough methodology to conduct a systematic review of PROMs (5). It consisted of a ten-step procedure to summarize the available evidence per measurement property per included PROM and draw conclusions on each measurement property per PROM. And subsequently, to give recommendations of the most suitable PROM for a given purpose, including also feasibility and interpretability aspects. This methodology also includes the COSMIN Risk of Bias checklist to assess the quality of studies on measurement properties of PROMs (1), including standards for design requirements and preferred statistical methods organized in boxes per measurement property.

To perform a systematic review on the quality of ClinROMs, PerFOMs and laboratory values, the same methodology can be used. However, we recommend some adaptations.

Two aspects of the COSMIN methodology for systematic reviews of PROMs are different for ClinROMs, PerFOMs or laboratory values: recommendation to use different boxes for reliability and measurement error, and the addition of a new step

The new boxes
In systematic reviews of ClinROMs, PerFOMs or laboratory values the COSMIN Risk of Bias checklist for PROMs (1) can be used, although the boxes for reliability and measurement error should be replaced with the COSMIN Risk of Bias tool to assess the quality of a study on reliability or measurement error (4). Standards for most of the remaining measurement properties (i.e. content validity, internal consistency, construct validity, criterion validity and responsiveness) developed for PROMs can be used for other types of measurement instruments as well. Some measurement properties are only relevant for multi-item instruments based on a reflective model (i.e. structural validity and internal consistency). For some other measurement properties only the final score or value of a measurement instrument is considered (i.e. hypotheses testing
for construct validity, criterion validity and responsiveness). The quality of studies on these measurement properties are similarly assessed for all types of outcome measurement instruments, and the existing boxes from the COSMIN Risk of Bias checklist for PROMs can be used.

An additional step
In a reliability study or a study on measurement error of a PROM the focus of interest is usually on the quality of the PROM as it is being used in clinical practice (analyzed using a one-way random effects model), or in the test-retest reliability (using a two-way random effects model of agreement). However, the focus of interest in a reliability study of other types of measurement instruments is much more diverse. As explained in chapter 2, there are many potential sources of variation (i.e. many different ways to operationalize the components of outcome measurement instruments) that could be the focus of interest in a study on reliability. Each result of all those studies tells you something about the quality of the instrument (and gives suggestions for improvement of the measurement by standardizing or restricting the source of variation which showed the largest error). Based on an overview of all these studies, an best-evidence measurement protocol can be recommended.

In a COSMIN reviews of ClinROMs, PerFOMs or laboratory values, an additional step is needed in the ten-step procedure (see Figure 3), specifically in the assessment of reliability and measurement error. To well interpret the results of studies included in a systematic review, you need to decide how the results of the study you want to assess inform you about the quality of the measurement instrument. Therefore, we separated the assessment of reliability and measurement error from the other measurement properties.

Change in the methodology
Based on our experience using the methodology, we decided to remove step 8 (which was ‘Evaluate interpretability and feasibility’) from the methodology. Aspects of interpretability and feasibility are only extracted (and summarized) rather than evaluated. Therefore, this step is irrelevant in the methodology. However, we consider it very useful to have a separate step on data extraction. Once you included all the studies in a review, we first recommend you to extract all necessary information from an article, before assessing the risk of bias, and the quality of the instrument. Relevant information to be extracted refers to characteristics of the included measurement instruments, information on feasibility and interpretability, characteristics of the studies, and the results of the study.

Consequently, the step-numbers are deviating from the step numbers presented in the original 10-step procedure of the COSMIN methodology to conduct a systematic review of PROMs (5).
Figure 3. Eleven-step procedure for conducting a systematic review on any type of outcome measurement instrument.

1. Formulate the aim of the review
2. Formulate eligibility criteria
3. Perform a literature search
4. Select abstracts and full-text articles
5. Extract data on characteristics of included measurement instruments, and information on feasibility and interpretability
6. Evaluate the content validity
7. Evaluate the internal structure
   - Structural validity
   - Internal consistency
   - Cross-cultural validity
8. Evaluate:
   - Reliability
   - Measurement error
9. Evaluate other measurement properties
   - Criterion validity
   - Hypotheses testing for construct validity
   - Responsiveness

Determine how the study result informs you on the quality of the instrument

Evaluate the quality of the instrument:
   - Extract data on studies and results of studies
   - Evaluate the methodological quality of the included studies by using the COSMIN boxes
   - Apply criteria for good measurement properties by using quality criteria
   - Summarize the evidence
   - Grade the quality of the evidence by using the GRADE approach

Only in step 8:
   - Draw conclusion on best-evidence measurement protocol

10. Formulate recommendations
11. Report the systematic review
4.1 The eleven-step procedure for conducting a systematic review of ClinROMs, PerFOMs, or laboratory values

Below, a summary is given for the eleven-step procedure. In the user manual of the COSMIN methodology for systematic reviews of PROMs (45) a thorough explanation of each step is provided. Only the steps that are different for a review of outcome measurement instruments other than PROMs are described here in detail. Please note that the number of the step are changed.

The methodology of a systematic review of outcome measurement instruments is subdivided into three parts (A, B, and C) (5).

**Step 1-4: Perform the literature search**

The steps 1-4 are standard procedures when performing systematic reviews, and are in agreement with existing guidelines for reviews (46, 47): formulating the specific aim of the review, and the eligibility criteria, performing the literature search, and selecting relevant publications.

In the research question, and eligibility criteria four key elements should be included: 1) the construct; 2) the population; 3) the type(s) of instruments; and 4) the measurement properties of interest.

In the search strategy we recommend to also use these key elements, except from the type of instruments, as we are not aware of highly sensitive search blocks for different types of measurement instruments. Search filters for different constructs may be found at https://blocks.bmi-online.nl/. When using the search filter for finding studies on measurement properties (48) of ClinROMs, PerFOMs and laboratory values, we recommend to use additional search terms for finding studies using Generalizability theory. This string, developed with the help of a clinical librarian, can be added with the boolean “OR” to the search filter.

**Pubmed search string for finding studies using Generalizability theory:**


**EMBASE search string for finding studies using Generalizability theory:**

‘g-theory’.ti:ab OR ‘g theory’:ti,ab OR ‘generalizability theory’:ti,ab OR ‘generalisability theory’:ti,ab
Step 5: Data extraction

Once you included all relevant articles, you check per article which measurement properties were evaluated (and subsequently decide which COSMIN boxes are relevant to be completed for the specific article). When reading through the article, at this point, we recommend you to extract all information from the article about the characteristics of the included measurement instruments (for suggestions of characteristics see appendix 4), including aspects of feasibility and interpretability (see below).

Interpretability is defined as the degree to which one can assign qualitative meaning (that is, clinical or commonly understood connotations) to a quantitative score or change in scores of an outcome measurement instrument (7). Both the interpretability of single scores and the interpretability of change scores is informative to report in a systematic review. The interpretation of single scores can be outlined by providing information on the distribution of scores in the study population or other relevant subgroups, as it may reveal clustering of scores, and it can indicate floor and ceiling effects. The interpretability of change scores can be enhanced by reporting M(C)IC values. However, there is an ongoing debate about how these values should be assessed.

Feasibility is defined as the ease of application of the measurement instrument in its intended context of use, given constraints such as time or money (49). Aspects of feasibility are, for example, completion time, cost of an instrument, length of the instrument, type and ease of administration. Feasibility applies to both the patients and the professional who are involved in the measurement. The concept ‘feasibility’ is related to the concept ‘clinical utility’, where feasibility refers to a measurement instrument, and clinical utility refers to an intervention (50).

Interpretability and feasibility are not measurement properties because they do not refer to the quality of an outcome measurement instrument. However, they are considered important aspects for a well-considered selection of an outcome measurement instrument.
**Steps 6-9: Evaluate the measurement properties**

The steps 6-9 concern the evaluation of the nine measurement properties of the included outcome measurement instruments. In these steps per measurement property, data is extracted on the characteristics of the studies, and the result of each study, the risk of bias of the included studies is rated by using the COSMIN Risk of Bias standards, and the results of the studies are rated by applying the criteria for good measurement properties. Subsequently, all evidence is summarized, and the quality of all available evidence per measurement property per measurement instrument is graded using a modified GRADE approach.

Characteristics of the studies refer to the characteristics of the included patient populations, and population of included professionals (for suggestions of characteristics see appendix 5). For specific recommendations for extracting information on the results of studies on reliability and measurement error see step 8 extracting information (p53).

In step 6 the content validity is assessed. In step 7 the internal structure (structural validity, internal consistency and cross-cultural validity\measurement invariance) is assessed. As the assessment of reliability and measurement error requires an additional step (i.e. understanding how the results of a study inform you about the reliability or measurement error of a outcome measurement instrument), these two measurement properties are now assessed in a separate step, i.e. step 8, apart from the assessment of the measurement properties criterion validity, hypotheses testing for construct validity, and responsiveness (i.e. step 9).

**Step 6. Evaluate content validity**

In step 6 content validity is evaluated. In the current standards and criteria for assessing content validity of PROMs (6) emphasize is put on the relevance, comprehensiveness, and comprehensibility of the PROM for the construct, target population, and intended context of use. In this assessment also the development of the PROM is considered, specifically, the item elicitation phase and the results from the pilot-testing phase. The assessment of content validity of other types of instruments may be different, and more research is needed to develop standards and criteria for other types of measurement instruments.

Assessing the content validity of measurement instruments that include multiple items – either based on a reflective or formative model – can heavily lean on the standards and criteria for PROMs. Only, because professionals are involved in the measurement, the three aspects of content validity (i.e. relevance, comprehensiveness, and comprehensibility) should be asked to the professionals. Depending on the construct of interest, these aspects could be asked to patients, too, for example for PerFOMs, or ClinROMs about symptoms or severity of conditions.
For the assessment of content validity of measurement instruments that exist of a single parameter (e.g. imaging-based parameters, or laboratory values), other aspects are likely more relevant. For example, you should judge whether it makes sense that the measurement instrument indeed measures the construct it purports to measure, based on theory and medical knowledge, and based on the claims by the manufacturer. In addition, the unit of measurement should match the construct to be measured. For example, a 6 minute walk test – expressed in the distance covered over a time of 6 minutes – measures walking capacity, rather than physical functioning (51). As currently no standards and criteria for content validity exist, face validity (which is a rather subjective judgment about whether the content of the instrument indeed looks as an adequate reflection of the construct to be measured) could be assessed by the reviewer.

**Step 7. Evaluate the internal structure**

In step 7 the internal structure (structural validity, internal consistency and cross-cultural validity/measurement invariance) is assessed. This step is only relevant when the measurement instrument is a multi-item instrument based on a reflective model. The standards (1) and criteria (5) provided for systematic reviews of PROMs can be used.

**Step 8. Evaluate reliability and measurement error**

Next, in step 8 reliability and measurement error are assessed. In chapter 2 and 3 we have explained how to assess the quality of each study on reliability and measurement error.

In a systematic review per study, you should first extract information about the elements of a comprehensive research question (see chapter 2), the specific ICC model or formula, and the results of each study. Next, you should assess the study quality using the standards (see chapter 3), and assess the results of each study, by comparing the results against the criteria for good measurement properties (5). Subsequently, you should summarize all evidence for reliability and for measurement error, respectively, and grade the quality of the evidence using the modified GRADE approach (5). Based on this overview, you can recommend on the best-evidence measurement protocol for a specific measurement instrument.

**Extracting information**

In Appendix 1 we provide an example of a data extraction table. First, we recommend to extract the seven elements of a comprehensive research question, and the research
question as stated by the authors in the article. Based on the elements, you can subsequently formulate a comprehensive research question. Next, we recommend to extract the information about the key elements of the review, i.e. the construct, population, type of measurement instrument, and measurement properties of interest. The construct to be measured (element 3 of a comprehensive research question), and the specific measurement properties (element 4 of a comprehensive research question) are already extracted, so the target population and the type of measurement instrument are recommended to be extracted. The target population refers to the target population of the specific study. In the example of Skeie et al. (19), the target population were patients with low-back pain. This can be different from the study population (i.e. the sample used) as extracted in item 7, or (slightly) different from the target population of the review (e.g. a broader population). In the study of Skeie, not only patients with low-back pain were included, but also patients with other spinal complaints such as mid back pain, neck pain, and/or extremity pain, or even pain-free subjects. The type of measurement instrument refers to whether the instrument under study is a ClinROM, PerFOM, laboratory value, a PROM or an ObsROM.

Last, we recommend to extract information about the statistics: the model or formula used, the result, and, if applicable, its 95% confidence interval. If available, we recommend to extract the variance components, or the SD_{sample} or SD_{difference} (see also chapter 3.2 for more explanation). For ordinal or dichotomous data we recommend to extract the raw numbers in the cells plus marginal totals.

_Risk of Bias assessment_

The next step in the review, is to assess the quality of each study, using Part B of the Risk of Bias tool to assess reliability and measurement error (as described in chapter 3). We recommend to use the worst-score counts methods to come to an overall rating per study. In Appendix 2 we provide an example of such a table to organize these ratings. We recommend that each study is assessed by two independent reviewers, and that they come to consensus.

_Comparison against the criteria for good measurement properties_

Each result of each single study on reliability or measurement error is now compared against the criteria for good measurement properties (5). As no criteria for the unweighted Kappa, and CV were provided in the guidelines for systematic reviews of PROMs, we added these missing criteria (see Table 8). Criteria for % specific agreement are difficult to set, because they are, just like sensitivity and specificity, highly dependent on the situation. As a rule of thumb 80% might be used.
Table 8. Extended criteria for good reliability and measurement error (adapted from Prinsen et al. (5))

<table>
<thead>
<tr>
<th>Reliability</th>
<th>+</th>
<th>ICC or (weighted) Kappa ≥ 0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>?</td>
<td>ICC or (weighted) Kappa not reported</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>ICC or (weighted) Kappa &lt; 0.70</td>
</tr>
<tr>
<td>Measurement error</td>
<td>+</td>
<td>SDC or LoA or CV*√2*1.96 &lt; M(C)IC1; % specific agreement &gt; 80%2</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>MIC not defined</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>SDC or LoA or CV*√2*1.96 &gt; M(C)IC1; % specific agreement &lt; 80%2</td>
</tr>
</tbody>
</table>

1 the M(C)IC value may come from another study. 2 Sometimes a higher percentage is more appropriate; when substantiated, this could be appropriate, too.

**Summarizing the evidence**

To come to an overall conclusion of the reliability or the measurement error of an outcome measurement instrument, one should first decide whether the results from multiple studies can be combined. You should take two aspects into account in this decision. 1) Do the results refer to the same information (i.e. refer to the same underlying comprehensive research question). Results from different designs (i.e. different components were varied across the repeated measurements) give you other information about the reliability of an instrument, and therefore cannot simply be summarized. And 2) Are the results consistent, that is all results are either sufficient (+) or insufficient (-). In case of inconsistency in results, we recommend to search for reasons for this inconsistency, e.g. different designs or statistical models, different populations, different background of raters. Subsequently, subgroups of studies can be summarized.

To summarize the evidence, you can either qualitatively summarize the results (e.g. describe the range of the results) or quantitatively pool the results. In reliability studies, only the point estimate of an ICC or Cohen’s kappa is used to conclude whether the specific measurement instrument has sufficient reliability (e.g. in the criteria that we propose above). Therefore, it is not necessary to pool the data to obtain a more precise point estimate.

The measurement error refers to the absolute deviation of the score from the ‘true’ score or the amount of error in the score. The point estimate of the measurement error parameter refers to this deviation or error, and therefore it is used to know how precise the measurement instrument is able to measure a patient. To come to a more precise point estimates of the measurement error, the parameters obtained in studies with the same design (i.e. that have the same underlying comprehensive research question) can
be pooled, when the confidence intervals are also reporting (which can be obtained using the sample size (39) or bootstrapping methods (52)).

Note, that you should only summarize or pool parameters of measurement error that were derived from the same study design and model or formula used. For example, the SEM\text{consistency} (either formula 4 or 8, chapter 3.2) and SEM\text{agreement} (formula 2, chapter 3.2) should not be combined. However, SEM\text{consistency} using either formula 4 or 6 (chapter 3.2) can be combined as they will lead to the same result, and the SDC\text{consistency} using either formula 5, 7, or 9 (chapter 3.2) can be combined. The same results are found when using either the SEM\text{one-way random effects model} (formula 1, chapter 3.2) or SEM\text{agreement} (formula 2, chapter 3.2). This is because all sources of variance (apart from the variance between patients) are taken into account in both formulas. Therefore, these parameters can be combined.

\textit{Handling inconsistent results.}

If the results of studies with the same underlying research question are inconsistent (e.g. both sufficient and insufficient results are found), first explanations for inconsistency should be explored. For example, slightly different study populations or methods were used. If an explanation is found, subgroups of studies (e.g. now based on the same study population, or in which the same source of variation is varied) can be summarized. The overall conclusion for (e.g.) reliability can subsequently be drawn per subgroup. When the explanation is found in the quality of the studies (i.e. very good and adequate studies lead to another overall rating than doubtful and inadequate studies), the doubtful and inadequate quality studies may only be reported, but ignored in this step, and only very good and adequate quality studies are considered to be decisive in determining the overall rating when ratings are inconsistent. This should be explained in the manuscript.

If studies with the same underlying research question showed inconsistent results, and no explanation can be found, one can conclude that results are inconsistent.

We refer to the User manual of the COSMIN methodology for systematic reviews of PROMs for more information.

\textit{Rate the quality of the summarized result}

If multiple studies can qualitatively be summarized (e.g. the range of results) or quantitatively pooled, the overall result can again be compared to the criteria for good measurement properties (see Table 8); you can then conclude that the outcome measurement instrument has either sufficient (+) or insufficient (-) reliability or measurement error. Or you should conclude that the results are inconsistent (±), or
indeterminate (?). For more information, we refer to the User manual of the COSMIN methodology for systematic reviews of PROMs.

**Grading the quality of the evidence using the modified GRADE approach**

After summarizing or pooling all evidence per outcome measurement instrument for reliability and for measurement error, and rating the summarized or pooled results against the criteria for good measurement properties, the next step is to grade the quality of the evidence. The quality of the evidence refers to the confidence that the summarized or pooled results is trustworthy. We developed a modified GRADE (Grading of Recommendations Assessment, Development, and Evaluation) approach to grade the evidence as high, moderate, low or very low (5), based on the 1) risk of bias (i.e. the methodological quality of the studies), 2) inconsistency (i.e. unexplained inconsistency of results across studies), 3) imprecision (i.e. total sample size of the available studies), and 4) indirectness (i.e. evidence from different populations than the population of interest in the review). This procedure is described in the User manual of the COSMIN methodology for systematic reviews of PROMs (5, 45).

**Draw conclusion on 'best-evidence measurement protocol’**

The results of reliability studies with their specific designs inform you whether a source of variation (for example the training of a rater, the specific machine used) importantly affects the score (i.e. the measurement). If possible, this source of variation should be standardized or restricted in future measurements. By looking at all evidence for various source of variation, you can now draw conclusions about how to standardize and restrict the measurement, and describe this best-evidence measurement protocol.

**Step 8. Evaluate criterion validity, hypotheses testing for construct validity, and responsiveness**

In step 8 criterion validity, hypotheses testing for construct validity, and responsiveness is assessed. The standards (1) and criteria (5) provided for systematic reviews of PROMs can be used.
Steps 10-11: Select the outcome measurement instrument

The steps 10 and 11 concerns the formulating recommendations (step 10) and the reporting of the systematic review (step 11).

Step 10. Formulate recommendations

The goal of a systematic review on measurement instruments is to get an overview of all available evidence on the quality of outcome measurement instruments that measure a specific construct in a defined patient population. Based on this overview, and taking aspects of feasibility and interpretability into account, we recommend you to formulate your recommendations about the most suitable outcome measurement instrument. To come to an evidence-based and fully-transparent recommendation, we recommend to categorize the included measurement instruments into three categories. Per type of measurement instrument you can conclude which instrument(s) are recommended (category A) or promising (category B), or insufficient (category C) and should not be used any more.

Category (A):

We recommend using different definitions of the category (A), depending on the structure of the measurement instrument:

<table>
<thead>
<tr>
<th>Multi-item reflectief</th>
<th>Evidence for sufficient content validity (any level), AND sufficient internal consistency (at least low quality, meaning also sufficient structural validity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-item formatief</td>
<td>Evidence for sufficient content validity (any level)</td>
</tr>
<tr>
<td>Single item (single parameter) (no gold standard)</td>
<td>Sufficient face validity (rated by e.g. the reviewers team), AND evidence for sufficient reliability (any level)</td>
</tr>
<tr>
<td>Single item (gold standard available)</td>
<td>Evidence for sufficient criterion validity, AND evidence for sufficient reliability (any level)</td>
</tr>
</tbody>
</table>

Category (B): outcome measurement instrument not categorized as 'A' or 'B'.

Category (C): outcome measurement instrument with high quality evidence for an insufficient measurement property.
Step 11. Report the systematic review

In accordance with the PRISMA Statement (53, 54), we recommend to report the following information:

(1) the search strategy (for example on a website or in the (online) supplemental materials to the article at issue), and the results of the literature search and selection of the studies and measurement instruments, displayed in the PRISMA flow diagram (including the final number of articles and the final number of measurement instruments included in the review) (Appendix 3);

(2) the characteristics of the included measurement instruments, including aspects of feasibility and interpretability (Appendix 4);

(3) the characteristics of the studies, including the characteristics of the included patient populations, and population of included professionals (Appendix 5);

(4) the methodological quality ratings of each study per measurement property per measurement instrument (i.e. very good, adequate, doubtful, inadequate), the results of each study, and the accompanying ratings of the results based on the criteria for good measurement properties (sufficient (+) / insufficient (-) / indeterminate (?)). In the User Manual for conducting systematic reviews of PROMs (45) an example is provided. In Appendix 6 we provide examples specifically for columns on reliability and measurement error. The table could be published for example as Appendix or supplemental material on the website of the journal only;

(5) a Summary of Findings (SoF) table per measurement property, including the pooled or summarized results of the measurement properties, its overall rating (i.e. sufficient (+) / insufficient (-) / inconsistent (±)/ indeterminate (?)), and the grading of the quality of evidence (high, moderate, low, very low). In the User Manual for conducting systematic reviews of PROMs (45) an example is provided. In Appendix 7 we provide examples specifically for columns on reliability and measurement error. These SoF tables (i.e. one per measurement property) will ultimately be used in providing recommendations for the selection of the most appropriate PROM for a given purpose or a particular context of use.
Appendix 1. Data Extraction table of relevant information for each included study in a systematic review.

<table>
<thead>
<tr>
<th>Extraction item</th>
<th>Instruction</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elements of a comprehensive research question</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Name of the instrument</td>
<td>Alternatively: type of instrument and parameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Version or way of operationalization</td>
<td>All relevant components that are known or expected to influence the score, and which are standardized or restricted (facet of stratification (23))</td>
<td>Equipment: Preparatory actions:</td>
<td>Equipment: Preparatory actions:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unprocessed data/sample collection:</td>
<td>Unprocessed data/sample collection:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data processing and storage:</td>
<td>Data processing and storage:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assignment of the score/determination of the value:</td>
<td>Assignment of the score/determination of the value:</td>
</tr>
<tr>
<td>3. Construct</td>
<td>Description of what is being measured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Measurement property</td>
<td>Reliability and/or measurement error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Components that will be repeated</td>
<td>e.g. whole measurement (i.e. all components) or some of the component</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Source(s) of variation varied</td>
<td>Components which is varied across the measurements (i.e. focus of analysis; facet of generalizability (23))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Patient population</td>
<td>(i.e. facet of differentiation (23))</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The research question</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Published research question</td>
<td>As formulated by the authors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehensive research question</td>
<td>As formulated by the reviewer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional key element of research aim of the review</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target population</td>
<td>Description of the population to which the authors want to generalize</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Types of measurement instrument</td>
<td>e.g. ClinROM, PerFOM, laboratory value, PROM or ObsROM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical information and results</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model or formula used</strong></td>
<td>Statistical model</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Result</strong></td>
<td>e.g. results (95% CI) of ICC, kappa, SEM, LoA and systematic difference</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variance components</strong></td>
<td>All reported variance components</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Apply criteria for good measurement property</strong>*</td>
<td>sufficient (+), insufficient (-), or indeterminate (?)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*although this is a rating, and not data extraction, we include it here, as the required information to make the rating is extracted here.*
Appendix 2. Risk of Bias ratings per standard per study

<table>
<thead>
<tr>
<th>Risk of Bias rating</th>
<th>study 1</th>
<th>rater 1</th>
<th>rater 2</th>
<th>consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design requirements</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Stability of the patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Time interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Similarity of measurement condition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Administration without knowledge of scores or values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Score assignment or determination of values without knowledge of the scores or values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Other important flaws</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Statistical methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 For continuous scores: ICC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 For ordinal scores: Kappa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 For dichotomous/nominal scores: Kappa for each category against the other categories combined?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Final rating</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3. Example of a Flow-chart

- PubMed
  n = ...

- EMBASE
  n = ...

- Additional record identified
  n = ...

Records screened after removing duplication
n = ...

Article selected based on title and abstract
n = ...

Reasons for exclusion:
- Validation not aim of the study (n=)
- Different construct measured (n=)
- Different study population (n=)
  Other (n=)

Total included in the review:
n articles describing x measurement instruments
Appendix 4. Example of reporting table on characteristics of the included measurement instruments.

<table>
<thead>
<tr>
<th>Name (reference to first article)</th>
<th>Construct</th>
<th>Intended context of use</th>
<th>Best-evidence measurement protocol</th>
<th>Target population</th>
<th>Type of measurement instrument</th>
<th>Feasibility aspects</th>
<th>Interpretability aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMM thickness (19)</td>
<td>Thickness of resting muscle</td>
<td>Evaluation</td>
<td>Training diagnostic ultrasound. Specific instructions for patient, and probe positions.</td>
<td>Patients with low back pain</td>
<td>Ultrasound</td>
<td></td>
<td>Mean score in mix of pain patients was 27.9mm (±3.2)</td>
</tr>
<tr>
<td>LMM contraction (19)</td>
<td>Comparison of the thickness of resting muscle with that of activated muscle</td>
<td>Evaluation</td>
<td>Training diagnostic ultrasound. Specific instructions for patient, and probe positions.</td>
<td>Patients with low back pain</td>
<td>Ultrasound</td>
<td></td>
<td>Mean score in mix of pain patients ranges 1.3mm (±1.7) – 3.5mm (±2.6)</td>
</tr>
</tbody>
</table>

Other characteristics which may be extracted are: conceptual model used, recommended by standardization initiatives, full copy available, fit for purpose (diagnostic, prognostic, evaluation).

Aspects of feasibility are, for example, completion time, licensing information and costs of an instrument, type and ease of administration. Feasibility applies to both the patients and the professional who are involved in the measurement. It may be considered to report this information in a separate Table.

Aspects of interpretability refer to 1) interpretability of single scores (e.g. information on the distribution of scores in study population or other relevant subgroups, and floor and ceiling effects), and 2) interpretability of change scores (i.e. M(C)IC values).
Appendix 5. Example of reporting table on characteristics of the study populations.

<table>
<thead>
<tr>
<th>Measurement instrument</th>
<th>Reference</th>
<th>Measurement property assessed</th>
<th>Patient population</th>
<th>Professional population</th>
<th>Response rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(19) Study 2</td>
<td>Reliability, measurement error</td>
<td>30, 47% female, age mean (SD) 37 (±12); LBP n=20; neck/midback pain n=5; extremity pain n=1; pain free n=4</td>
<td>2, Chiropractors experienced in diagnostic ultrasound for the musculoskeletal system, i.e. 4 and 8 years resp., with a postgraduate diploma in diagnostic ultrasound. Before the study, both developed the protocol of diagnostic ultrasound that was applied in this study.</td>
<td></td>
</tr>
<tr>
<td>LMM contraction</td>
<td>(19) Study 3</td>
<td>Reliability</td>
<td>30, 50% female, age mean (SD) 38 (±11); LBP n=23; neck/midback pain n=7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(19) Study 4</td>
<td>Reliability, measurement error</td>
<td>30, 43% female, age mean (SD) 40 (±11); LBP n=20; neck/midback pain n=6; extremity pain n=3; pain free n=1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Patient characteristics refer to, e.g. age, gender, disease characteristics (diagnosis, disease duration, disease severity), setting, and geographical location.

Rater characteristics may refer to, e.g. professional background, specific training received, or years of experience.
Appendix 6. Overview Table of quality and results of studies on reliability and measurement error.

<table>
<thead>
<tr>
<th>Measurement instrument (MI) [ref]</th>
<th>Type of MI</th>
<th>Reliability</th>
<th>Measurement error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>Study quality</td>
</tr>
<tr>
<td>LLM contraction score (study 2)</td>
<td>Ultrasound</td>
<td>30</td>
<td>Adequate</td>
</tr>
<tr>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLM contraction score (study 3)</td>
<td>Ultrasound</td>
<td>30</td>
<td>Adequate</td>
</tr>
<tr>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLM contraction score (study 4)</td>
<td>Ultrasound</td>
<td>30</td>
<td>Adequate</td>
</tr>
<tr>
<td>(19)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLM contraction score (ref)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLM contraction score (ref)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled or summary result (overall rating)</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a calculated from LoA*
Appendix 7. Summary of Findings Tables for Reliability and Measurement error.
Based on the studies on reliability described by Skeie (19)

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Summary result</th>
<th>Overall rating</th>
<th>Quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrasound measurement of the LMM contraction score – best-evidence measurement protocol: rater, day and active motor tasks performed before measurement were not of influence</td>
<td>Range ICC: 0.94-0.97</td>
<td>Sufficient</td>
<td>High (two studies of adequate quality)</td>
</tr>
<tr>
<td>Measurement instrument B –</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on the studies on measurement error described by Skeie (19)

<table>
<thead>
<tr>
<th>Measurement error</th>
<th>Summary result</th>
<th>Overall rating</th>
<th>Quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultrasound measurement of the LMM contraction score – best-evidence measurement protocol: rater, day and active motor tasks performed before measurement were not of influence</td>
<td>Range SDC_{consistency} : 1.08 - 1.29 ( \text{MIC = not assessed} )</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Measurement instrument B –</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


42. de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. BMJ. 2013;346:f2125.
43. de Vet HC, Dikmans RE, Eekhout I. Specific agreement on dichotomous outcomes can be calculated for more than two raters. J Clin Epidemiol. 2017.
44. de Vet HCW, Mullender MG, Eekhout I. Specific agreement on ordinal and multiple nominal outcomes can be calculated for more than two raters. J Clin Epidemiol. 2018;96:47-53.