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

Version 1.0 dated December 2020

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

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

Cite:

Lidwine B. Mokkink, Maarten Boers, CPM van der Vleuten, LM Bouter, Jordi Alonso, Donald L 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).

Funding:

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).

Content

The COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error (in short: COSMIN Risk of Bias tool) consists of two parts: in Part A you (as the user of the tool) specify the comprehensive research question, in order to exactly understand the study; in Part B you assess the quality of the study.

To complete Part A, the user can disentangle the 7 elements of a comprehensive research question, based on the stated research question in the article and the design of the study.

To complete Part B, you as a user can assess the quality using 9 standards for a reliability study, or 8 standards for a study on the measurement error. Using the worst-score-counts principle, the quality of a study can be assessed.

A detailed user manual is published to help you using the tool.

PART A. Understanding HOW the study informs you on the reliability and measurement error of an outcome measurement instrument

When using this COSMIN Risk of Bias tool, the user should first understand how exactly the result informs him/her about the research question. To fully understand this, the 7 elements of the research question should be clear. Based on what is described in the paper (e.g. the stated research question, but above all, the methods and design of the study), these 7 elements can be specified by the user of the COSMIN risk of Bias tool.

The tables with components of the outcome measurement instruments (see pages 6-12) may help you to think of with components were varied over the repeated measurements (i.e. the focus of the study) and which components were standardized, or (kept or assumed to be) stable. We developed two sets of components, separated for outcome measurement instruments that do and those that do not involve biological samples.

ELEMENTS OF A COMPREHENSIVE RESEARCH QUESTION

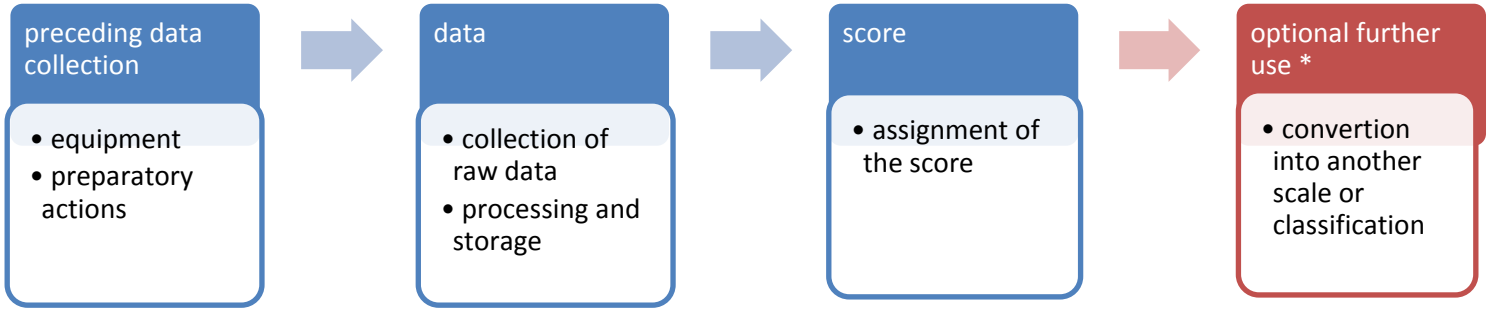
EXTRACT THE 7 ELEMENTS FROM THE STUDY TO UNDERSTAND HOW THE STUDY INFORMS YOU ON THE RELIABILITY OR MEASUREMENT ERROR OF THE OUTCOME MEASUREMENT INSTRUMENT

Table 1. Elements that make up a comprehensive reached question.

Element of the research question	
1	the **name** of the outcome measurement instrument
2	the **version** of the outcome measurement instrument or way of **operationalization** of the measurement protocol
3	the **construct** measured by the measurement instrument
4	a specification whether one is interested in a **reliability parameter** (i.e. a relative parameter such as an ICC, Generalizability coefficient ϕ, or Kappa κ) or a **parameter of measurement error** (i.e. an absolute parameter expressed in the unit of measurement e.g. SEM, LoA or SDC; or expressed as agreement or misclassification, e.g. the percentage specific agreement).
5	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)
6	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)
7	a specification of the **patient population** studied

ICC = Intraclass correlation coefficient; SEM = standard error of measurement; LoA = Limits of Agreement; SDC = smallest detectable change.

Components of outcome measurement instruments that do not involve biological sampling



* Not part of the outcome measurement instrument; further (optional) use or interpretation of the score is done by a linear or semi-quantitative conversion into another scale or classification

Figure 1. Components (i.e. potential sources of variance) of outcome measurement instruments that do not involve biological sampling

The left three boxes of the figure are considered to make up the outcome measurement instrument (or measurement protocol). The right side box reflects further (optional) *use or interpretation* of the outcome measurement instrument. Optionally, the (continuous) value can be converted into another scale or classification, for example, into an ordinal scale (e.g. none/ mild/ moderate/ severe) or dichotomous scale (e.g. below or above a normal value), or into a responder on treatment (e.g. below or above the Minimal Important Change (MIC) value). As this is a linear or semi-quantitative conversion, it is not considered to be a potential source of variance.

Table 2. Components of outcome measurement instruments that do not involve biological sampling

Component	**Elaboration**	**Examples**
Equipment	All equipment necessary in the preparation, the administration, and the assignment of scores of the outcome measurement instrument	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.
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[1] • preparations of the patient[2] by the professional • Preparations undertaken by the patients	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. Listen to and understanding the instructions provided; adherence to the preparatory instructions such as fasting, resting, taking medication, bowel preparation, exercising, shaving.

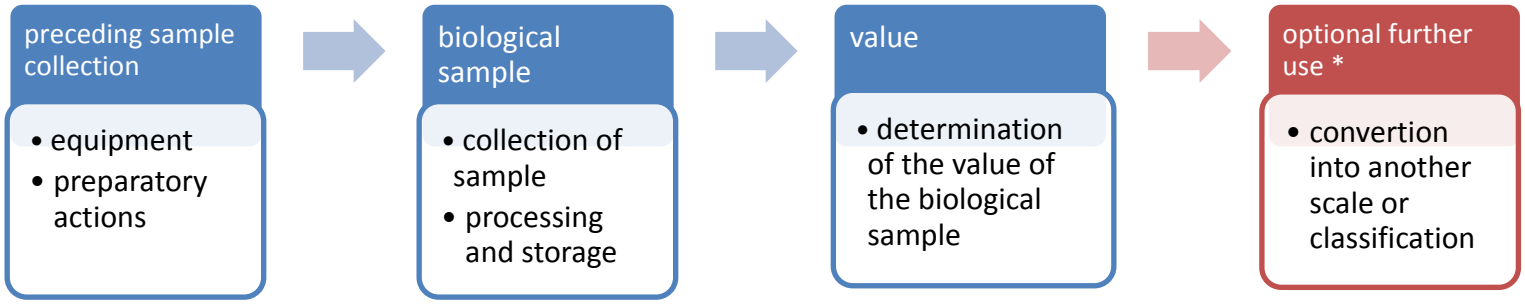
Component	Elaboration	Examples
Collection of raw data	All actions undertaken by patient and professional(s) to collect the data, before any data processing	The patient completing questions at home, or at the hospital; or performing the tasks; the rater observing or timing the performance; switching the imaging device on and off; positioning and moving the ultrasound probe.
Data processing and storage	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)	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.
Assignment of the score(s)	Methods used to convert processed data into a score[3] that constitutes the outcome measurement instrument.	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 or control area of an image; scores adjusted for e.g. missing data or patients using devices such as mobility aids.

[1] 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.

[2] In the COSMIN methodology we use the word 'patient.' However, sometimes the target population is not patients, but e.g. healthy individuals, caregivers, or clinicians, or body structures (e.g. joint, or lesion). In these cases, the word patient should be read as e.g. healthy volunteer, clinician, or the relevant body structure.

[3] The score can be further used or interpreted, by converting a score to another scale, metric or classification. However, this is not part of the outcome measurement instrument. 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).

Components of outcome measurement instruments that involve biological sampling



* Not part of the outcome measurement instrument; further (optional) use or interpretation of the value is done by a linear or semi-quantitative conversion into another scale or classification

Figure 2. Components (i.e. potential sources of variance) of outcome measurement instruments that involve biological sampling

Table 3P. Components (i.e. potential sources of variance) of outcome measurement instruments that involve biological sampling

Component	**Elaboration**	**Examples**
Equipment	All equipment used in the preparation, the administration, and the determination of the values of the outcome measurement instrument	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.
Preparatory actions preceding sample collection by professionals, patients, and others (if applicable)	1. General preparatory actions, such as required expertise or training for professionals to prepare, administer, store and determine the value	Training, education or experience required, certification.
	2. Specific preparatory actions for each measurement, such as • preparations of equipment, environment, and storage by professionals[1]	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.
	• preparation of the patient[2] by the professional	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.
	• Preparatory actions undertaken by the patients	Listen to and understanding the instructions provided; adherence to the preparatory instructions such as fasting, resting, taking medication, exercising, shaving, washing of hands.

Component	Elaboration	Examples
Collection of biological sample	All actions undertaken to collect the biological sample, before any sample processing	Taking a blood sample or tissue biopsy, collection of a sample of urine 'mid-stream' in a container.
Biological sampling processing and storage	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	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 tetrasodium salt of ethylenediaminetetraacetic 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.
Determination *of the value* of the biological sample	Methods used to count or quantify the amount of the substance or entity of interest[3]	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.

[1] 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.

[2] 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. joint, or lesion). In these cases, the word patient should be read as e.g. healthy volunteer, clinician, or the relevant body structure.

[3] 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.

PART B. Assessing the quality of a study on reliability or measurement error

The COSMIN Risk of Bias tool contains standards that can be used to determine whether the result of an study can be trusted. To assess the quality of a study, each standard should be rated, and the worst-score-count method will be applied (standards which are not applicable will not be included in determining the final rating) to determine the risk of bias.

The Box Reliability contains 6 standards about design requirements, and three standards on the preferred statistical methods for studies on reliability. The Box Measurement Error contains the same six standards about the design requirements, as the design for studies on reliability and for studies on measurement error is the same, and the same data can be used for estimating both measurement properties. Next, the Box Measurement Error contains two standards about the preferred statistical methods for studies on measurement error.

Standards for studies on reliability

	Design requirements	very good	adequate	doubtful	inadequate	NA
1	Were patients stable in the time between the repeated measurements on the construct to be measured?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	Na
2	Was the time interval between the repeated measurements appropriate?	Yes		Doubtful, OR time interval not stated	No	Na
3	Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Yes (evidence provided)	Reasons to assume standard was met, OR change was unavoidable	Unclear	No (evidence provided)	Na
4	Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
5	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?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
6	Were there any other important flaws in the design or statistical methods of the study?	No		Minor methodological flaws	Yes	

Statistical methods		**very good**	**adequate**	**doubtful**	**inadequate**
7	For continuous scores: was an intraclass correlation coefficient (ICC) calculated?	ICC calculated; the model or formula was described, and matches study design and the data	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	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	
8	For ordinal scores: was a (weighted) kappa calculated?	Kappa calculated; the weighting scheme was described, and matches the study design and the data	Kappa calculated, but weighting scheme not described or does not optimally match the study design		
9	For dichotomous/nominal scores: was Kappa calculated for each category against the other categories combined?	Kappa calculated for each category against the other categories combined			

Standards for studies on measurement error

	Design requirements	very good	adequate	doubtful	inadequate	NA
1	Were patients stable in the time between the repeated measurements on the construct to be measured?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	Na
2	Was the time interval between the repeated measurements appropriate?	Yes		Doubtful, OR time interval not stated	No	Na
3	Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Yes (evidence provided)	Reasons to assume standard was met, OR change was unavoidable	Unclear	No (evidence provided)	Na
4	Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
5	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?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
6	Were there any other important flaws in the design or statistical methods of the study?	No		Minor methodological flaws	Yes	

Statistical methods	**very good**	**adequate**	**doubtful**	**inadequate**
7 For continuous scores: was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC), Limits of Agreement (LoA) or Coefficient of Variation (CV) calculated?	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	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	$SEM_{consistency}$ $SDC_{consistency}$ or LoA or CV calculated, without knowledge about systematic difference or with evidence provided that systematic difference has occurred	SEM calculated based on Cronbach's alpha, or using SD from another population
8 For dichotomous/nominal/ordinal scores: Was the percentage specific (e.g. positive and negative) agreement calculated?	% specific agreement calculated	% agreement calculated		