OMERACT Instrument Selection

Topic: Critical Appraisal

This document provides readers with a guide to various resources on critical appraisal using OMERACT Instrument Selection methodology.

A. Guidance to critical appraisal

A.1. Instrument selection overview whiteboard:

https://omeract.org/instrument-selection/ [see 6:20]]

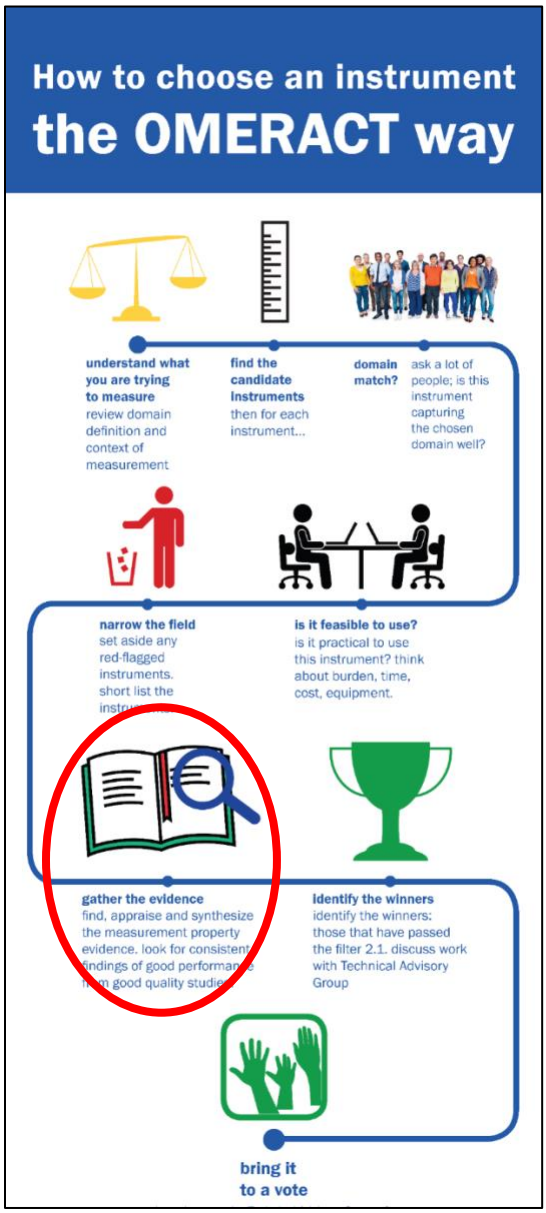
A.2. Critical appraisal video:

https://omeract.org/instrument-selection/

A.3. Instrument selection detailed discussion video:

https://omeract.org/instrument-selection/ [see 19:22]

B. OMERACT Way

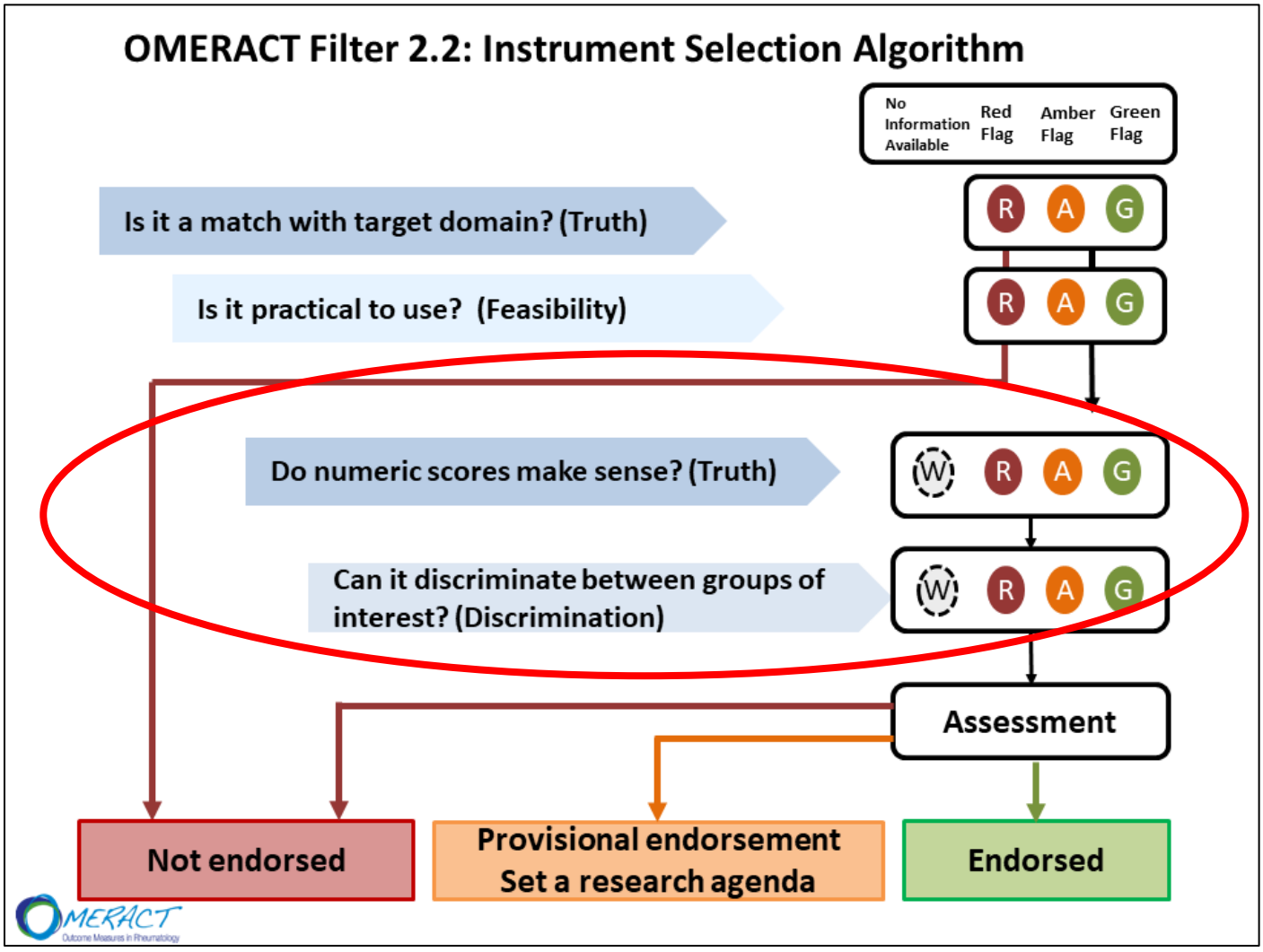


C. OMERACT Master checklist for instrument selection. Step 9: Critical appraisal

OMERACT Master Checklist for Instrument Selection ***Name of Instrument*:**		
Step #	**OMERACT Instrument Selection Process Checklist Item**	Mark when complete
Assembly of working group and protocol development		
1	Assemble working group	o
2	Decide on methods protocol for Core Outcome Instrument Set selection	o
3	**Deliverable**: Submit protocol using Instrument Selection Workbook to Technical Advisory Group [TAG]	o
4	Review and approval of final protocol by TAG	o
Review of evidence of instrument performance for existing or new instrument		
Part A: Domain match and Feasibility assessment		
5	Obtain Working Group and others assessment of match with the target domain	o
6	Obtain Working Group and others assessment of feasibility	o
7	Is the instrument a match with the domain AND feasible? Yes ____ → if yes, continue with Part B of checklist below No ____ → If no, set instrument aside (find new one or develop new one)	o
Part B: Review of evidence of performance of an instrument across key measurement properties		
8	Conduct literature search; create PRISMA diagram; place articles of measurement properties in Summary of Measurement Properties (SOMP) Table	o
9	Conduct COSMIN-OMERACT Good Methods check, add findings into the SOMP Table	o
10	Conduct data extraction, create summary reporting tables, fill in SOMP Table with assessment of adequacy of results	o
11	Conduct synthesis across evidence available for each measurement property	o
12	Decide if any gaps exist in evidence of measurement properties If gaps found, draft protocol for new study to fill gaps If no gaps, finish the SOMP Table with proposed level of endorsement	o
Initial submission to TAG: literature review findings & protocol for gaps		
13	**Deliverable**: Submit the Instrument Selection Workbook to TAG	o
14	Receive final response from TAG	o
15	If studies are needed to fill gaps, conduct new measurement property studies, submit to TAG for Good Methods check, add to body of evidence (SOMP) and go back to Step 12 If no studies are needed, put X here: _____and move to Step 16	o
Final submission to TAG for approval		
16	Obtain agreement on final report	o
17	Set timeline for next review of instrument	o
Ratification of level of endorsement by OMERACT Community and communication of results		
18	Ratification of level of endorsement by OMERACT Community	o
19	Implement communication and dissemination plan	o

D. OMERACT Filter 2.2. Instrument Selection Algorithm (OFISA)

Each study contributing evidence to the questions 'Do numeric scores make sense? (Truth)' and 'Can it discriminate between groups of interest? (Discrimination)' are assessed for good quality methods using the COSMIN-OMERACT Good Methods Check.



E. Where does critical appraisal fit on the SOMP?

The critical appraisal of each study using the COSMIN-OMERACT Good Methods Check is shown by the colours in each cell in the SOMP.

Green = Good methods used – use this evidence

Amber = Some cautions, but this will be used as evidence

Red = There are some problems – do not use this evidence

Instrument: ABC Domain: Physical function					Date completed: 2021-02-11			
Population: rheumatoid arthritis	Intervention(s): drug		Control: placebo/drug		Type of studies: clinical trials			
Author/year	Truth Domain match	Feasibility	Truth		Discrimination			
			Construct validity	Inter-method reliability	Test retest reliability	Long'l construct validity	Clinical trial discrimination	Thresholds of meaning
Working Group Appraisal (n=20 including 7 PRPs)	+	+						
Tugwell 2005			+/–			+		
Shea 2004						+		+
Smith 1999								
Beaton 2015							+	
De Wit 2018							+	
Wells 2004			+					
March 2008							+	+/–
D'Agostino 2011						+/–		+
Bingham 2018			+		+/–			
Singh 2010			+					
Strand 2015			+/–					
Simon 2011						+		+/–
New data from Conaghan 2021					+			
Total available studies for each property			5	N/A	3	5	3	4
Total studies available for synthesis			5	N/A	2	4	3	4
Synthesis Rating	**GREEN From Working group**	**GREEN From Working group**	**GREEN**	**N/A**	**AMBER**	**GREEN**	**GREEN**	**AMBER**
OMERACT Endorsement	**Based on the OMERACT algorithm this instrument is: Provisionally endorsed *More research needed on test-retest reliability and thresholds of meaning.***							

F. Excerpt from OMERACT Handbook, Chapter 5, Instrument Selection (pg. 37-39)
https://omeracthandbook.org/handbook

9. Conduct COSMIN-OMERACT Good Methods check, add findings into the SOMP Table

The X's on the Summary of Measurement Properties table for each measurement property show the pool of potential evidence that is for each measurement property (i.e. you can see the total available studies for each property). However, some studies may have flaws in their methods that make them at risk for misestimating the true value for the measurement property. Whiting (2011) suggest biases occur when "systematic flaws or limitations in the design or conduct of a study distort the results" (Whiting 2011, pg. 529). Pieces of evidence like these should be excluded from the review. This is the same as a risk of bias assessment in other types of systematic reviews. There are many tools available to critically appraise the methods used in measurement studies, but few have a focus on this risk of bias that we needed. One instrument, the popular COSMIN (Consensus-based Standards for the selection of health Measurement Instruments) methodological quality appraisal checklist (Mokkink 2010; Terwee 2012) did discuss features of a study that could, according to their expert panel and core working group, represent a risk of bias. In the COSMIN checklist these are the "POOR" or "INADEQUATE" ratings only. In 2015, in collaboration with its developers, we developed a modification of the COSMIN system, focusing on what would become the COSMIN Version 2.12 (Mokkink 2018) checklist as the source. In this 4-point methodological rating system, some COSMIN Version 2.12 items offer an "INADEQUATE" rating (in some versions a POOR rating). They offer this rating to only those items which the COSMIN group felt would indicate a methodological flaw that would warrant exclusion from evidence synthesis due to a risk of bias. Only a subset of COSMIN Version 2.12 items offer this rating and OMERACT has focused on this subset (Beaton 2019).

We assembled those items offering an INADEQUATE rating into a checklist and reworded and reversed each to be an affirmative statement. An affirmation of these would suggest avoidance of this particular risk of bias and therefore suggest that the study had used at least ADEQUATE or "good enough" quality of methods. Our approach therefore focuses only on avoiding those critical flaws in design and methods (risks of biasing the results) that would cause us to set aside this piece of evidence. This is consistent with the meaning of an inadequate score in the COSMIN approach. Importantly, we recognize that this depends on reported methods, rather than actual ones. Reported methods are usually used, given the difficulty in reaching primary authors of each measurement study. However, if groups do wish to contact the authors, this would be an evaluation of actual methods, and each set of authors would need to be contacted in order to be systematic in approach. We believe that as reporting standards begin to appear for measurement studies, there will be more congruence between reported methods and the critical features of the actual methods used. For now, we need to critique based on reported methods, recognizing that this does not necessarily mean the investigators overlooked things, rather they did not report on them.

Reviewers assess each study and give a rating of whether the article did critical good method (YES) or did not report doing it in their study (NO). Based on the array of YES and NO responses (and knowing that a NO would normally reflect an inadequate rating and a piece of evidence that would not be considered in the synthesis step), the reviewer makes a summary appraisal of whether, given the results of the Good Methods Check, this piece of evidence is trustworthy enough to be included. The checklist and the appraisal together are called the COSMIN-OMERACT Good Methods Check. Table 2 below shows one example for test-retest reliability.

Table 2. COSMIN-OMERACT Good Methods Check for Test-retest reliability. In this system (as is the case in COSMIN v2.12), a "No" or "Red" rating would indicate a serious methodological flaw that would suggest this piece of evidence should not	**Notes: (please keep notes about your ratings, and**

be considered. In the COSMIN-OMERACT Good Methods Check, the reviewer then makes an overall decision about inclusion or exclusion of this evidence.			**your final decision).**
	Yes, good methods	No, not done well	
Were patients stable in the interim period on the construct to be measured?			
Was the time interval appropriate?			
Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions			
Were the statistical methods appropriate (choose one from below)? • A. For continuous scores: Was an intraclass correlation coefficient (ICC), Pearson correlation or Spearman correlation calculated? • B. For dichotomous (yes/no) ordinal or nominal scores (named but not ordered categories: red hair/brown hair/blond hair): Was kappa calculated?			
Otherwise good methods? (Free of any other important flaws in design or methods).			
Considering the information available, would you recommend this study as evidence to be considered for this measurement property? (enter this in Summary of Measurement Properties) |__| Yes, good methods used – use this evidence |__| Some cautions, but this will be used as evidence |__| No, there are some problems – do not use this evidence.			
Notes on this piece of evidence:			

There were no fatal flaw checklists available in COSMIN for two of the OMERACT Filter 2.2 measurement properties (thresholds of meaning and sensitivity to changes in clinical trial settings) for which we created our own list based on critical elements in their design as discussed in the literature (Beaton 2011; Bossuyt 2003; Higgins 2011; Schmitt 2015; Whiting 2004; Whiting 2011). Devji et al. have since published an assessment of the credibility of anchor-based methods that has been integrated into the thresholds of meaning quality appraisal (Devji 2021).

It is recommended that two independent reviewers complete the Good Methods Check and then check for consensus. All ratings and the final the Good Methods consensus vote should be kept for the records and will be part of the work submitted to the TAG of OMERACT at the end of this process. The instrument workbook has the good methods check table for each measurement property and there is an Excel spreadsheet available to working groups to track this evaluation. The overall consensus will be entered into the Summary of Measurement Properties Table using the colours GREEN [for good methods], AMBER [some caution but consensus this evidence should go forward] or RED [for problematic methods and an indication that this study will not be used in synthesis]. Look back at the Summary of Measurement Properties table in Figure 5.7 and see that the cells are coloured in for the example studies.

Remember that each article could address more than one measurement property. If a concern is found about the risk of bias related to one property, that evidence is excluded. However, the next good methods check on the next property could show that very good methods were used for it, and that evidence will continue to be used.

9	Conduct COSMIN-OMERACT Good Methods check, add findings into the SOMP Table	○

G. Excerpt from Instrument selection workbook (pg. 27-32)
https://omeracthandbook.org/workbooks

9. Check to see if each of the included studies has used good methods when assessing each measurement property using the COSMIN-OMERACT good methods check; add these findings into the SOMP Table by coloring cells Green, Amber, or Red.

9.1 COSMIN-OMERACT Good Methods Check

Once you have your articles and their measurement properties organized, you then need to do a "COSMIN-OMERACT Good Methods Check" (i.e. a quality appraisal) on the methods used to evaluate each measurement property in each article. Good Methods should be checked by two raters and agreement reached. After the checks have been done, an overall rating is given by the pair of raters to say whether they feel this piece of evidence should go forward for further assessment of the adequacy of the results.

Below are the **COSMIN-OMERACT Good Methods Checklists** *for each of the measurement properties in the OMERACT Filter 2.2. Use one table per study; e.g. if you found 3 studies assessing construct validity, you will need 3 of the tables below. In order to help you track your Good Methods Checklist results, we have created a spreadsheet* (LINK TO EXCEL WORKSHEET) *based on work by Alessandro Chiarotto who kindly shared his template for us to adapt based on the following reference: Chiarotto A, et al. Measurement properties of Numeric Rating Scale, Visual Analogue Scale and Pain Severity subscale of the Brief Pain Inventory in patients with low back pain, a systematic review. J Pain. 2019 Mar;20(3):245-263.*

You can use either the Word tables below or the Excel spreadsheet to report the Good Methods Check results.

Pillar: TRUTH **Question: Do the numeric scores make sense?** **Measurement property: Construct (hypothesis testing) validity (COSMIN Space 8)**			
Author Year	Yes, good methods used	No, not achieved	Notes
Was a clear description given of the construct measured by the comparator instrument?	☐	☐	Click here to enter text.
Were the measurement properties of the comparator instrument(s) described and at least adequate?	☐	☐	Click here to enter text.
Were design and statistical methods adequate for the hypotheses to be tested?	☐	☐	Click here to enter text.
Otherwise good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.
Considering the information available, would you recommend this study as evidence to be considered for this measurement property? *(enter this in the OMERACT Summary of Measurement Properties Table)*			

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

Pillar: TRUTH
Question: Do the numeric scores make sense?
Measurement property: Inter-method reliability (e.g. inter-rater, inter-machine)

Author Year	Yes, good methods used	No, not achieved	Notes
Were the measurements conducted independently?	☐	☐	Click here to enter text.
Did the design of the study hold all other factors constant except for the source of variability being examined?	☐	☐	Click here to enter text.
Were the test conditions similar for the measurements? (e.g., type of administration, environment, instructions)	☐	☐	Click here to enter text.
Was the correct statistic used? • Continuous data: intra-class correlation coefficient (ICC) used. • Dichotomous/ordinal/nominal scores: Kappa (w) used.	☐	☐	Click here to enter text.
Otherwise good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.

Considering the information available, would you recommend this study as evidence to be considered for this measurement property? *(enter this in the OMERACT Summary of Measurement Properties Table)*

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

Pillar: DISCRIMINATION
Question: Can it discriminate between situations of interest?
Measurement property: Test-retest reliability (COSMIN Space 5)

Author Year	Yes, good methods used	No, not achieved	Notes
Were the patients stable in the interim time period?	☐	☐	Click here to enter text.
Was the time interval appropriate?	☐	☐	Click here to enter text.

Were the test conditions similar for the measurements? (e.g., type of administration, environment, instructions)	☐	☐	Click here to enter text.
Was the correct statistic used? • Continuous data: intra-class correlation coefficient (ICC) used. • Dichotomous/ordinal/nominal scores: Kappa used.	☐	☐	Click here to enter text.
Otherwise good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.

Considering the information available, would you recommend this study as evidence to be considered for this measurement property? *(enter this in the OMERACT Summary of Measurement Properties Table)*

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

Pillar: DISCRIMINATION
Question: Can it discriminate between situations of interest?
Measurement property: Responsiveness (Longitudinal Construct validity) (COSMIN Space 9 a,b,d)

Author Year	Yes, good methods used	No, not achieved	Notes
Can the criterion for change be considered an adequate gold standard OR is the construct for change clear (either as a situation of change or an actual indicator of change)?	☐	☐	Click here to enter text.
Were the measurement properties of the comparator standard described and at least adequate? (N/A for "gold standards).	☐	☐	Click here to enter text.
Were the statistical methods appropriate for the testing situations? (for comparison to gold standard this would include ROC, AUC, predictive values, sensitivity & specificity; correlation of change with external anchor, for constructs: effect size, standardized response mean, correlation).	☐	☐	Click here to enter text.
Otherwise good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.

Considering the information available, would you recommend this study as evidence to be considered for this measurement property? (*enter this in the OMERACT Summary of Measurement Properties Table*)

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

Pillar: DISCRIMINATION
Question: Can it discriminate between situations of interest?
Measurement property: Clinical trial discrimination (COSMIN Space 9c)

Author Year	Yes, good methods used	No, not achieved	Notes
Was the time interval between testing stated and appropriate?	☐	☐	Click here to enter text.
Were there a proportion of people expected to change in one or both groups? (Improvement or deterioration)?	☐	☐	Click here to enter text.
Were hypotheses formulated regarding the anticipated mean differences in change scores between subgroups a priori? • i.e. positive/negative or no change can be expected.	☐	☐	Click here to enter text.
Were the statistical methods adequate for the hypotheses tested (relative efficiencies, pooled treatment effect sizes, standardized mean differences)?	☐	☐	Click here to enter text.
Otherwise good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.

Considering the information available, would you recommend this study as evidence to be considered for this measurement property? (*enter this in the OMERACT Summary of Measurement Properties Table*)

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

Pillar: DISCRIMINATION
Question: Can it discriminate between situations of interest?
Measurement property: Thresholds of meaning

Author Year	Yes, good methods used	No, not achieved	Notes
Was the patient group similar to your target population (level of disease severity, demographics)?	☐	☐	Click here to enter text.
Is the anchor easily understandable?			
Is the anchor clearly related to the target domain of interest (i.e. good correlation between anchor and instrument)?			
Was the cut-off on the anchor used to MID justified to be a small but important difference/important state?			

Did the same respondent respond to instrument and anchor?			
Was analysis done separately for improvement and deterioration OR only in same direction anticipated in the target application?	☐	☐	Click here to enter text.
Were multiple criteria and/or analyses used and results triangulated?	☐	☐	Click here to enter text.
Did the analysis include either a Youden index threshold from ROC, or another cut off on an ROC approach? Or if a threshold type of approach (25% or 75%) was used, was it tested for diagnostic utility (sensitivity and specificity)?	☐	☐	Click here to enter text.
Otherwise, good methods? (Free of any other important flaws).	☐	☐	Click here to enter text.

Considering the information available, would you recommend this study as evidence to be considered for this measurement property? (*enter this in the OMERACT Summary of Measurement Properties Table*)

☐ **Yes, likely low risk of bias.**
☐ **Some cautions, but this will be used as evidence**
☐ **No, don't use this evidence**

In this spreadsheet you can use colour to track the responses of each rater to the Good Methods Checklist items.

Sequential columns show other articles included in this review (same as the rows on the OMERACT Summary of Measurement Properties Evidence Table). An example of one measurement property is listed below.

	COSMIN-OMERACT GOOD METHODS CHECKLIST			
	Article:	AUTHOR YEAR		
	Instruments:			
	Rater:	reviewer 1	reviewer 2	CONSENSUS
	Yes, likely low risk of bias			
	Some cautions, but this will be used as evidence			
	No, don't use this as evidence			
	Construct (hypothesis testing) validity			
1	Was a clear description given of the construct measured by the comparator instrument?	Yes, good methods	Yes, good methods	Yes, good methods
2	Were the measurement properties of the comparator instrument(s) described and at least adequate?	Yes, good methods	No, not achieved	Yes, good methods
3	Were design and statistical methods adequate for the hypotheses to be tested?	Yes, good methods	Yes, good methods	Yes, good methods
4	Otherwise good methods? (Free of any other important flaws)	Yes, good methods	Yes, good methods	Yes, good methods
	DECISION: LIKELY LOW RISK OF BIAS; SOME CAUTIONS; DON'T USE THIS EVIDENCE	Yes, likely low risk of bias	Some cautions, but this will be used as evidence	Yes, likely low risk of bias

9.2 Fill in SOMP with results of Good Methods Check

Colour the cells in the SOMP with the result of each assessment of the Good Methods Check, either GREEN, AMBER, or RED.