Instrument Selection Methods sessions @OMERACT 2022

Session: Adequacy of results

Goal/Objective of instrument methods sessions:

1. Improve understanding of topic

2. Engage TAG members with topic as part of their training

3. Provide platform for working groups to present their experience of using the methodology

4. Obtain advice on improving strategies for using this methodology

Background information on adequacy of results

1. Instrument selection overview whiteboard:

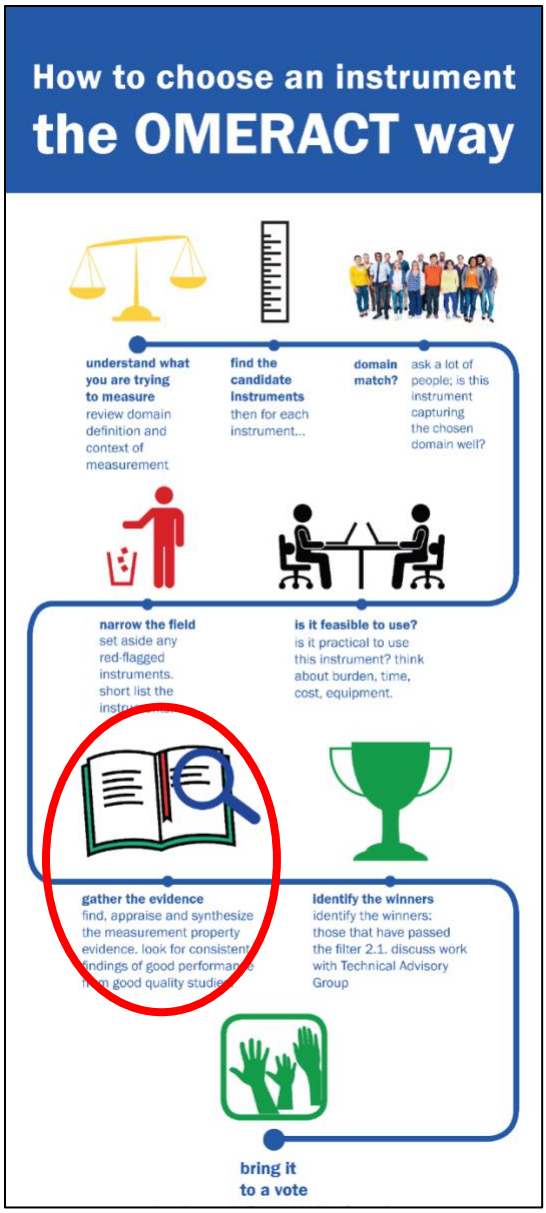
https://omeract.org/instrument-selection/ [see 6:38]

2. Instrument selection detailed discussion video:

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

(note this video is going to be updated but the information for judging adequacy of results is still relevant)

2. OMERACT Way

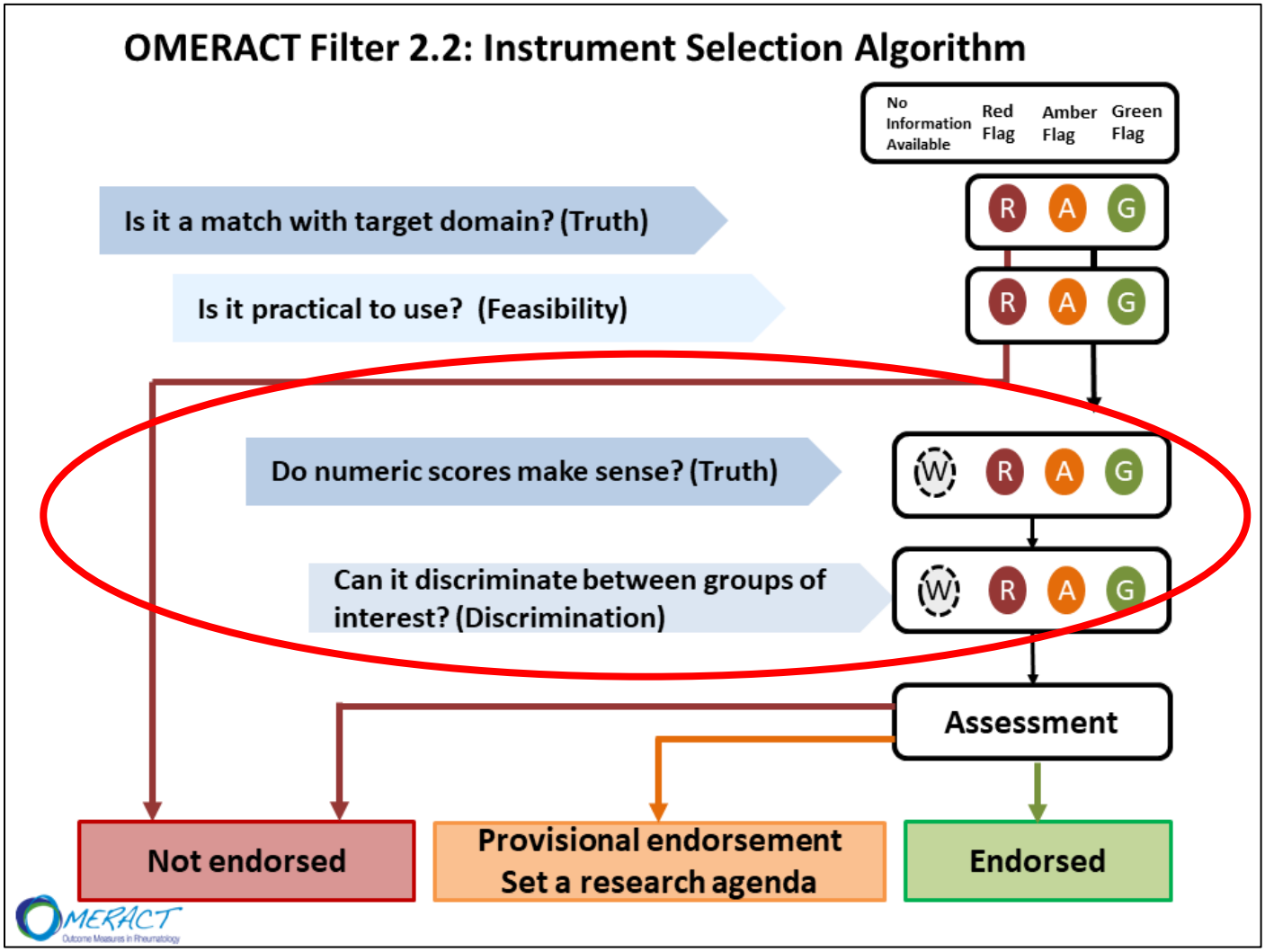


3. Master checklist for instrument selection: Step 10

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

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

Each study judged as Green or Amber using the COSMIN-OMERACT Good Methods Check will then have its results extracted and judged to see if they meet the OMERACT provisional adequacy standards.



5. Excerpt from OMERACT Handbook, Chapter 5, Instrument Selection (pg. 39-41)

10. Conduct data extraction, create summary reporting tables, fill in SOMP Table with assessment of the adequacy of results

10.1 Data extraction and completion of summary reporting tables for each measurement property

Studies that have passed the COSMIN-OMERACT Good Methods Check with either a Green or an Amber rating are now reviewed to extract information and the results of the measurement property tests. Information that we extract from the study includes descriptive information on the study, study population and methods, and the results for each measurement property. These results are then compared to standards for acceptable evidence of validity or reliability.

What information should be extracted? The OMERACT TAG has created templates for the reporting of each measurement property guided by existing suggestions of the key elements for each (Beaton 2001; Lohr et al., 1996; McLeod et al., 2011; Mokkink et al., 2010a; Mokkink et al., 2010b; Mokkink et al., 2009; Nunnally and Bernstein, 1994; Nunnally and Durham, 1975; Reeve et al., 2013; Scientific Advisory Committee of the Medical Outcomes, 2002; Terwee et al., 2012; Valderas et al., 2008; Wyrwich et al., 2013) as well as literature related to that measurement property itself. In addition, details about each study (e.g. population description, sample size, study design/methods) should also be extracted using the general study description template.

At a minimum, data extraction should be done in duplicate to pilot test accuracy in at least five articles or 5% of articles. The working group can decide on the level of accuracy and whether a single reviewer can do remaining data extraction or if it should continue to be done in duplicate. This decision and the results of the pilot should be documented. If possible, we would strongly recommend duplicate extraction of information on the measurement studies.

The tables should enable the extraction of the most important features of the study to allow the most useful and useable information to be available to future readers. These features overlap with those that could create a risk of bias, so these tables also provide information readers can use to appreciate the quality of the methods used in the study. The tables can be long and should be done thoroughly from the outset to prevent having to go back into articles for more information in the future.

The data extraction templates are available in Appendix C and the instrument selection workbook also links to these tables.

10.2 Comparing findings to published standards

Defining the standards to indicate when a study is demonstrating that an instrument has good enough reliability and validity is highly variable in the literature. The OMERACT TAG undertook a review of these standards and has developed from them a provisional set of standards (for at least adequate performance) for each property. Adequacy is *only* examined in the evidence that has been determined to have "good methods" (green or amber in cell). The provisional standards are included in Table 3 below. A full description of the standards we reviewed in order to come to these thresholds is available from the TAG. Standards will be ratified at a future OMERACT and their "provisional" nature removed.

Table 3: OMERACT Filter 2.2 Provisional adequacy standards

Pillar (and Question)	Measurement property	OMERACT Filter 2.2 Provisional standards for adequate performance
Truth. (***Question 3***. Do the numeric scores make sense?)	Internal consistency	Not part of Filter 2.2, if included should be alpha >0.75, higher if target application is individual clinical decision making (0.90).
	Construct validity	Pre-specified hypotheses are replicated. Should be shown with similar constructs, dissimilar constructs and known groups in order to show both presence and absence of a relationship as appropriate.
	Inter-method reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered adequate for a Green rating) Excellent needed for measurement if done for individual clinical decision making. Please also report on SEdiff and MDD-95, Bland-Altman graph is helpful.
Discrimination (***Question 4***: Can it discriminate between groups of interest?)	Test retest reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered adequate for a Green rating) Excellent needed for measurement if done for individual clinical decision making. Please also report on SEdiff and MDC-95, Bland-Altman graph is helpful.
	Longitudinal construct validity	Consistency with a priori theory in studies that look at situation similar to the intended application. Anticipated large effect expect SRM >0.80, medium/moderate effect, SRM 0.5-0.79, small effect 0.2-0.5. Findings outside the anticipated range should be considered a negative finding.
	Sensitivity in clinical trials	Longitudinal data are provided for the groups that have changed and separately for groups that have remained stable or had a different amount of change compared to the first group. SRM is greater in change group than in stable or different change group. This difference is also reported in a relative effectiveness statistic ($ES_{group1}^2/ES_{group2}^2$) = hypothesized magnitude and direction.
	Thresholds of meaning	There are not "standards" for a calculated threshold. We ask only that reporting and context be as clear as possible for users. Report threshold value and how it was calculated, error boundaries if possible. Thresholds should be related to the anchors used (i.e., threshold for predicting disease activity), sensitivity and specificity of the cut point. For change thresholds, describe relation of both MID and MDC and guide interpretation accordingly.

Working Groups use a + sign to indicate that that piece of evidence exceeds the provisional standard for that property, a – sign when it does not meet that standard, and a +/- for inconsistent findings (for example in testing construct validity several comparisons could be made). These symbols are added to the Summary of Measurement Properties table in the respective slot.

10	Conduct data extraction, create summary reporting table, fill in SOMP table with assessment of the adequacy of results	○

6. Excerpt from Instrument selection workbook (pg. 32-33)

10. Conduct data extraction on those measurement properties that were assessed as green or amber Good Methods and complete measurement property tables for summary descriptions of the studies; fill in SOMP Table with assessment of the adequacy of results

10.1 Conduct data extraction and summary tables for each measurement property

The Working Group now needs to create tables to summarize the study characteristics and the actual findings of your evidence. Only measurement property assessments that have passed the COSMIN-OMERACT Good Methods Checklist with AMBER or GREEN ratings will be included. TAG has drafted summary tables to report each measurement property. We have also drafted a table where a general description of each included study can be reported. Groups may choose to format their own tables, but we ask that all the elements in the tables below are included. This includes the study design elements, as well as the analytic approach and results.

For current versions of summary tables for each of the measurement property tables, please click here: **https://omeract.org/instrument-selection/downloadable-forms/**

10.2 Judging the PERFORMANCE of the instrument based on the results found in the studies.

Below are the OMERACT provisional standards for adequate performance. Use this to guide your decisions to complete the judgement of the adequacy section in the summary tables. We use the following symbols:

+ = positive support for that measurement property.

+/- = ambivalent support, inconclusive result.

- = support that this instrument did not reach performance standards for that property.

When your data extraction on the results of the assessment of the performance of the measurement property is complete, you will fill in the results in your SOMP Table. Change the X to a symbol that summarizes the results using the (+, +/-, -) symbols.

Copy the completed summary tables into this section of the workbook or submit the completed spreadsheet if you choose to use Excel to record your data.

Pillar (and Question)	Measurement property	OMERACT Filter 2.2 Provisional standards for adequate performance
Truth. (***Question 3***. Do the numeric scores make sense?)	Internal consistency	Not part of Filter 2.2, if included should be alpha >0.75, higher if target application is individual clinical decision making (0.90).
	Construct validity	Pre-specified hypotheses are replicated. Should be shown with similar constructs, dissimilar constructs and known groups in order to show both presence and absence of a relationship as appropriate.
	Inter-method reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered adequate for a Green rating) Excellent needed for measurement if done for individual clinical decision making. Please also report on SEdiff and minimal detectable difference (MDD)-95, Bland-Altman graph is helpful.
Discrimination (***Question 4***: Can it discriminate between groups of interest?)	Test retest reliability	Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw) Excellent > 0.90. Good >0.75 (considered adequate for a Green rating) Excellent needed for measurement if done for individual clinical decision making. Please also report on SEdiff and MDD-95, Bland-Altman graph is helpful.
	Longitudinal construct validity	Consistency with a priori theory in studies that look at situation similar to the intended application. Anticipated large effect expect SRM >0.80, medium/moderate effect, SRM 0.5-0.79, small effect 0.2-0.5. Findings outside the anticipated range should be considered a negative finding.
	Sensitivity in clinical trials	Longitudinal data are provided for the groups that have changed and separately for groups that have remained stable or had a different amount of change compared to the first group. SRM is greater in change group than in stable or different change group. This difference is also reported in a relative effectiveness statistic ($ES_{group1}^2/ES_{group2}^2$) = hypothesized magnitude and direction. If reporting on % exceeding a threshold of meaning, please use an empirical cumulative distribution function for each group and highlight the location of your thresholds of meaning.
	Thresholds of meaning	There are not “standards” for a calculated threshold. We ask only that reporting and context be as clear as possible for users. Report threshold value and how it was calculated, error boundaries if possible. Thresholds should be related to the anchors used (i.e., threshold for predicting disease activity), sensitivity and specificity of the cut point. For change thresholds, describe relation of both minimal important difference (MID) and minimal detectable change (MDC) and guide interpretation accordingly.

7. Where does adequacy of results fit on the SOMP?

The adequacy of the result of each study is judged according to the following:

+ = positive support for that measurement property.
+/- = ambivalent support, inconclusive result.
- = support that this instrument did not reach adequate performance standards for that property.

The symbol is then entered into the cell for each study. Note that the adequacy of the result is not judged for those studies that were judged as 'RED – do not use this evidence' in the Good Methods Check.

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.***							