OMERACT

Outcome Measures in Rheumatology

THE OMERACT HANDBOOK

FOR ESTABLISHING AND IMPLEMENTING CORE OUTCOMES IN CLINICAL TRIALS ACROSS THE SPECTRUM OF RHEUMATOLOGIC CONDITIONS

VERSION 2.1 | June 2, 2021

OMERACT

Outcome Measures in Rheumatology

Striving to improve endpoint outcome measurement through a data driven, iterative consensus process involving relevant stakeholder groups.

Contributors: Dorcas Beaton, Lara Maxwell, Shawna Grosskleg, Bev Shea, Peter Tugwell, Clifton O. Bingham III, Philip G. Conaghan, Maria-Antonietta D'Agostino, Catherine Hofstetter, Lyn March, Lee S. Simon, Jasvinder A Singh, Vibeke Strand, George Wells. With thanks to the many OMERACT participants, patients and other colleagues who contributed to the content of this Handbooks

Contact: OMERACT Secretariat 43 Bruyère St. Annex E Ottawa, ON K1N 5C7 Canada ▪ admin@omeract.org

Table of Contents

Chapter 5. Instrument selection for Core Outcome Measurement Sets

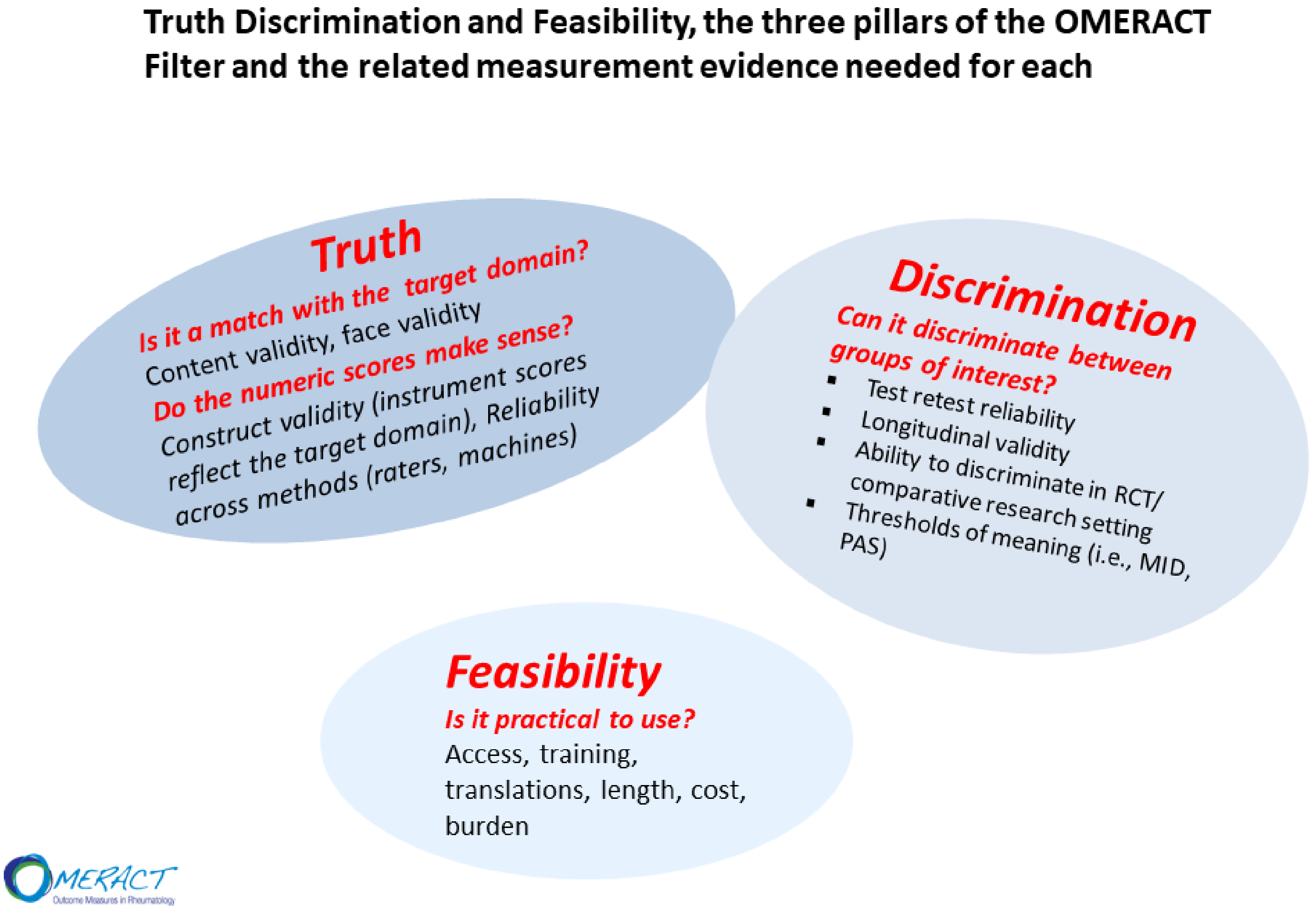
Instrument selection: Three pillars, four questions, one answer

Introduction and Background to Instrument Selection

OMERACT (Outcome Measures in Rheumatology) is an international, multi-stakeholder organization aiming to provide an evidence-based decision making process for agreement on patient-centred and important outcomes for use across clinical trials and observational studies in rheumatology (Boers et al., 1998; Boers et al., 2014; Boers et al., 2014a; Tugwell et al., 1993, Tugwell et al., 2007; Tugwell et al., 2014).

Since its inception in 1992, OMERACT has worked towards establishing Core Outcome Sets, that is, a minimum set of outcomes that should be measured in all clinical trials to allow for consistency and communication between trials. At OMERACT we divide the process into two phases, first developing a "Core Domain Set" (i.e. "what to measure" in clinical trials) and then embarking on a "Core Outcome Measurement Set" (i.e. "how to measure"). Completion of both leads to a Core Outcome Set.

Core Domain Sets and Core Outcome Measurement Sets are established through evidence- and consensus-based decision making across key stakeholder groups: Patients and caregivers, Providers, Payers, Product makers, Policy makers, Principal investigators (researchers) the Public, and others (e.g., the Press) (Tunis et al., 2017). The evidence that is needed is embodied in what OMERACT has called the "OMERACT Filter" which is made up of three pillars of evidence to ensure an instrument is fit for the purpose of use in a Core Outcome Measurement Set in clinical trials in a given disease group or field. The three pillars are: Truth, Discrimination, and Feasibility (See Figure 5.1). Outcome measurement instruments with sufficient evidence of these three pillars of the Filter were considered having the evidence to support their inclusion in a Core Outcome Measurement Set and to have passed the OMERACT Filter (Beaton et al., 2019; Beaton et al., 2020, Boers et al., 1998; Boers, Kirwan, Wells, et al., 2014; Boers et al. 2019; Maxwell et al., 2019). By 'measurement instrument', we mean a tool that is used to measure a quality or quantity of a variable. As defined in Boers et al, 2014, this "tool may be a single question, a questionnaire, a score obtained through physical examination, a laboratory measurement, a score obtained through observation of an image, and so on". An 'outcome measurement instrument' is a measurement instrument chosen to assess a specific outcome. Throughout the rest of this chapter, we will use the word 'instrument' to mean a 'measurement instrument' in this broader context.



Truth Discrimination and Feasibility, the three pillars of the OMERACT Filter and the related measurement evidence needed for each

Truth

Is it a match with the target domain?

Content validity, face validity

Do the numeric scores make sense?

Construct validity (instrument scores reflect the target domain), Reliability across methods (raters, machines)

Discrimination

Can it discriminate between groups of interest?

- Test retest reliability
- Longitudinal validity
- Ability to discriminate in RCT/ comparative research setting
- Thresholds of meaning (i.e., MID, PAS)

Feasibility

Is it practical to use?

Access, training, translations, length, cost, burden

Figure 5.1. The three pillars of the OMERACT Filter 2.2

A Core Outcome Measurement Set does not limit the investigator in choosing the primary outcome from that set, or in fielding other outcomes in their study. It advocates that each study should measure **at least** the Core Outcome Measurement Set. Consistent use of Core Domain Sets and Core Outcome Measurement Sets is increasingly recognized as important in maximizing comparability of findings across trials and facilitating meta-analyses and comparative effectiveness research (Beaton et al., 2015; Higgins et al., 2020; Williamson et al., 2012). They also reduce the risk of selective outcome reporting bias in clinical research. Kirkham has shown that the availability of the RA Core Set has increased consistency in outcome measurement in arthritis where in 2010 70% of trials utilized this core set in their outcomes (Kirkham et al., 2013).

Revision of the OMERACT Filter

In 2014, OMERACT began a deliberate process to refresh the guiding framework for domain and instrument selection (Boers, Idzerda, et al., 2014; Kirwan et al., 2014; Tugwell et al., 2014) . The revisions related to domain selection have been described (Boers, Kirwan, et al., 2014; Maxwell et al., 2019) with further elaboration of the OMERACT Filter in 2019 (Boers et al., 2019) and in Chapter 4 of this Handbook. The revisions in each text highlight the new framework and the need to have at least one domain represented from each of three Areas (Manifestations/Abnormalities, Life Impact and Death/Lifespan) and one strongly recommended one (Societal/Resource Use). Adverse events and important contextual factors are further decided upon and become part of each Core Domain Set. Once Core Domain Sets are endorsed by OMERACT (see Chapter 4 in this Handbook), Working Groups move on to identify at least one instrument that passes the Filter requirements for inclusion in a 'Core Outcome Measurement Set'. In response to the growing number of articles on measurement properties, and the growing number of instruments in the field, the OMERACT Filter 2.2 Instrument Selection process was also revised to help with the finding, appraising and synthesizing available or new evidence of measurement properties to ascertain the one answer that is sought – has the instrument passed the OMERACT Filter 2.2? In January 2017, the OMERACT Executive endorsed the process described in this chapter for instrument selection and in 2019 OMERACT published a manuscript describing this decision-making process entitled OMERACT Filter 2.1 (Beaton et al., 2019). In this version of the OMERACT Filter, 2.2, we reflect the changes made to ensure the Filter captures the type of evidence required for imaging outcomes. This further improved the process for all types of instruments allowing us to move forward with a consolidated single filter. We will monitor its use by working groups and seek their feedback to address areas for improvement.

The OMERACT Filter 2.2 instrument selection process has two functions:

1. To define the type of information that is needed to ascertain if an instrument has passed the Filter,
2. To suggest a process and provide tools to facilitate moving through Filter requirements and to facilitate record keeping and reporting.

To ensure transparency and rigour, Working Groups need to document the process they used and work towards a final report describing the body of evidence. A workbook and specific assistive tools have been developed to support the use of Filter 2.2 for instrument selection and to help track progress and findings (see Appendix A for the instrument selection workbook and the OMERACT website for instructional videos: https://omeract.org/instrument-selection/). Each instrument under consideration will have its own workbook. While the instrument selection process initially had an emphasis on patient-reported outcomes, the application to imaging outcomes and things like pulmonary function testing (Roofeh et al., 2021) and joint counts (Duarte-Garcia et al., 2019) add to our confidence that the revisions of the Filter can be applied across different types of clinical outcome assessments.

Foundation: How do we know if an instrument has passed Filter 2.2?

The original OMERACT pillars of Truth, Discrimination, and Feasibility remain the core pillars for instrument selection in OMERACT Filter 2.2.

In order to streamline the process of instrument selection, Filter 2.2 has suggested a practical re-ordering of the Filter elements (see Figure 5.2). The pillars of Truth, Discrimination, and Feasibility are still there, but are now ordered with each step reflecting an increasing investment of time and effort. It also suggests that after the first two steps, a decision can be made to stop considering that instrument before entering the most time-consuming part of the process in the literature review and creation of evidence that is needed for the last two elements.

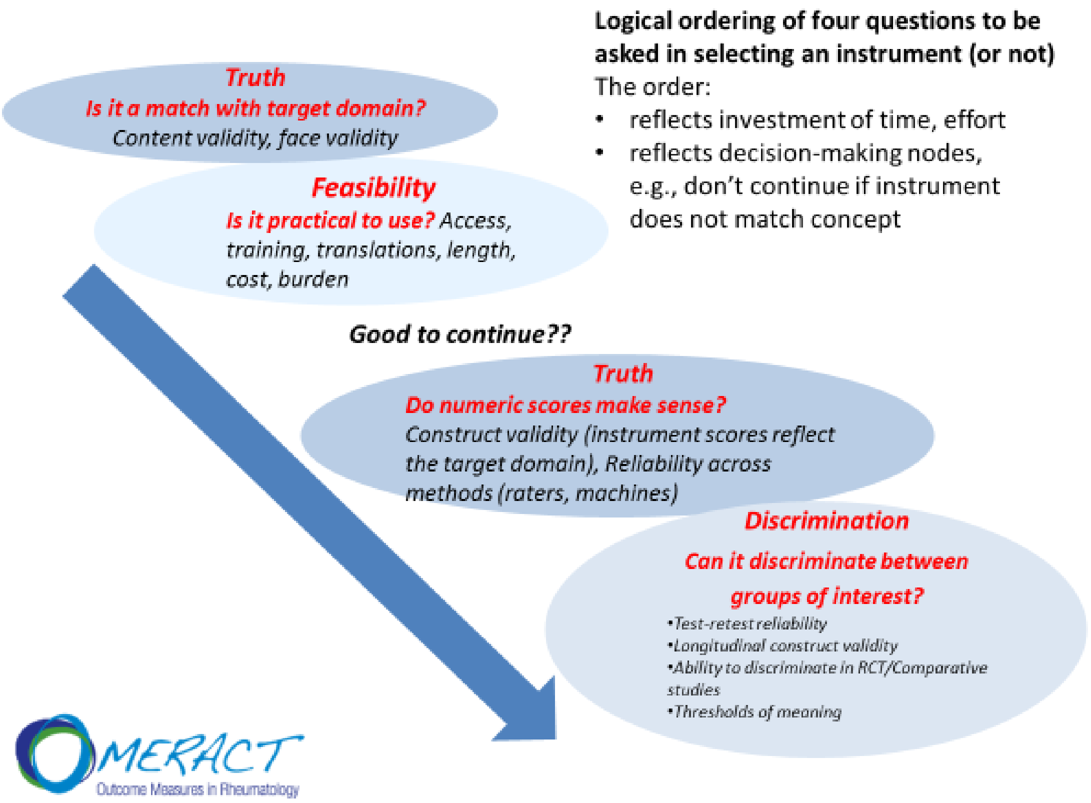
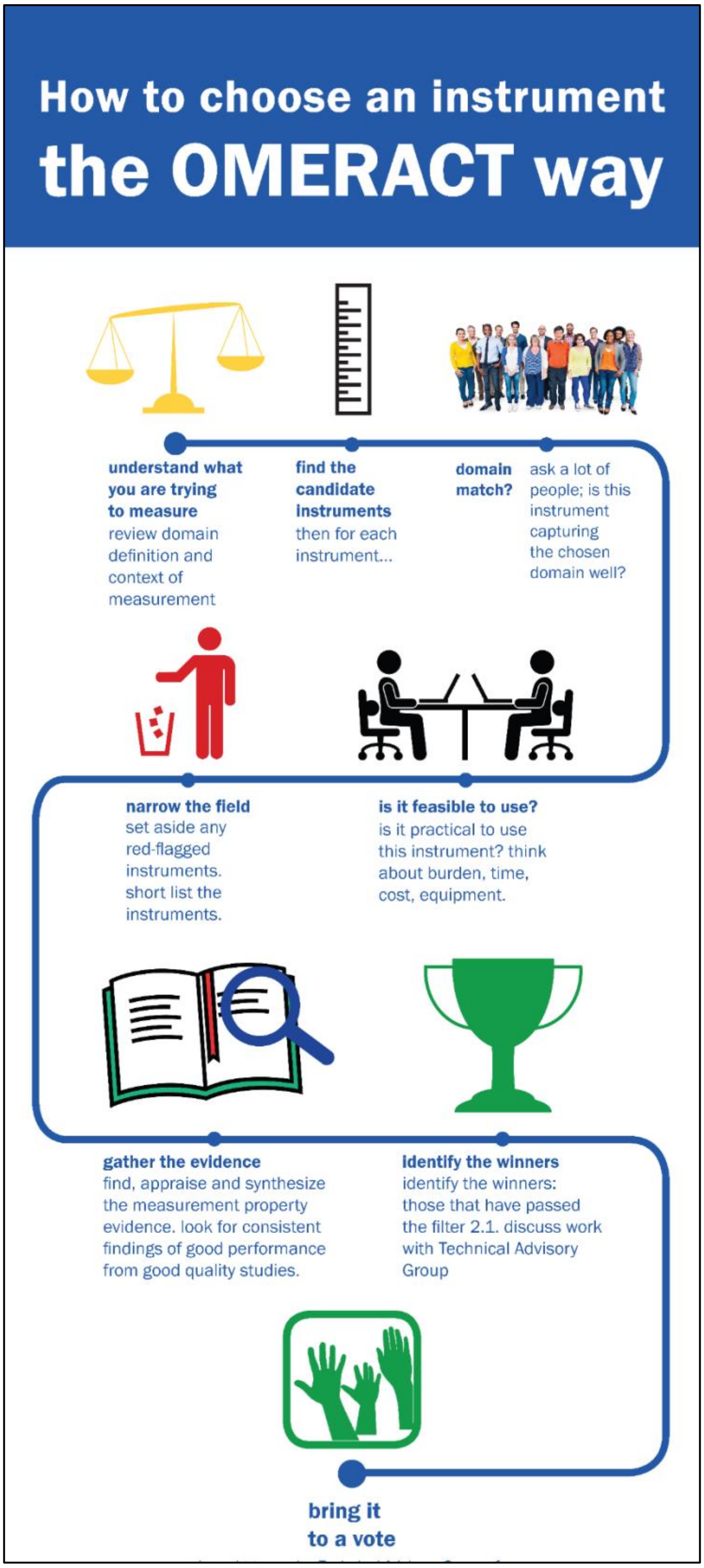

Logical ordering of four questions to be asked in selecting an instrument (or not)
The order:
• reflects investment of time, effort
• reflects decision-making nodes, e.g., don't continue if instrument does not match concept
Truth
Is it a match with target domain?
Content validity, face validity
Feasibility
Is it practical to use? Access, training, translations, length, cost, burden
Good to continue??
Truth
Do numeric scores make sense?
Construct validity (instrument scores reflect the target domain), Reliability across methods (raters, machines)
Discrimination
Can it discriminate between groups of interest?
•Test-retest reliability
•Longitudinal construct validity
•Ability to discriminate in RCT/Comparative studies
•Thresholds of meaning
OMERACT
Outcome Measures in Rheumatology

Figure 5.2 Re-ordering elements of Truth, Discrimination, and Feasibility to allow for logical, effort-saving decision-making process

This is when Working Groups may decide to let an instrument drop from consideration because it is not feasible, or upon closer inspection, it is decided that it is not a good match for their target domain. Only those instruments passing this decision point would go through the process of gathering evidence or developing new evidence to address Truth and Discrimination.

For scoring at each stage of the OMERACT Filter Instrument Selection Algorithm (OFISA), we have used a stoplight set of colours to reflect the Working Group's appraisal of whether an instrument met the Filter requirements for that attribute. GREEN is used to reflect a high level of confidence that this has been passed and can go forward. AMBER indicates that the group sees a need for additional work or has some concern, but still considers it good enough to pass that part of the Filter. RED indicates stop or does not pass. We also added the option of WHITE as an indicator of when no evidence is available (something that is correctable but must be corrected before the instrument has passed the Filter). WHITE is only offered for questions which are dependent on available evidence. **Again, GREEN and AMBER get a pass, RED and WHITE do not.**

We have created a graphical depiction of the overall process of Instrument Selection known as the 'OMERACT Way for Instrument Selection' (Figure 5.3). We have also created a 10-minute whiteboard video that provides an overview of this process: https://youtu.be/ym2n_bnRqP8



How to choose an instrument the OMERACT way

understand what you are trying to measure review domain definition and context of measurement

find the candidate instruments then for each instrument...

domain match? ask a lot of people; is this instrument capturing the chosen domain well?

narrow the field set aside any red-flagged instruments. short list the instruments.

is it feasible to use? is it practical to use this instrument? think about burden, time, cost, equipment.

gather the evidence find, appraise and synthesize the measurement property evidence. look for consistent findings of good performance from good quality studies.

identify the winners identify the winners: those that have passed the filter 2.1. discuss work with Technical Advisory Group

bring it to a vote

Figure 5.3. The OMERACT Way flowchart describing the step by step process

The OMERACT Master Checklist for Instrument Selection provides a step-by-step overview of the process for working groups to follow. We will now work through each step in the process.

	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	○
2	Decide on methods protocol for Core Outcome Instrument Set selection	○
3	**Deliverable**: Submit protocol using Instrument Selection Workbook to Technical Advisory Group [TAG]	○
4	Review and approval of final protocol by TAG	○
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	○
6	Obtain Working Group and others assessment of feasibility	○
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)	○
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	○
9	Conduct COSMIN-OMERACT Good Methods check, add findings into the SOMP Table	○
10	Conduct data extraction, create summary reporting tables, fill in SOMP Table with assessment of adequacy of results	○
11	Conduct synthesis across evidence available for each measurement property	○
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	○
Initial submission to TAG: literature review findings & protocol for gaps		
13	**Deliverable**: Submit the Instrument Selection Workbook to TAG	○
14	Receive final response from TAG	○
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	○
Final submission to TAG for approval		
16	Obtain agreement on final report	○
17	Set timeline for next review of instrument	○
Ratification of level of endorsement by OMERACT Community and communication of results		
18	Ratification of level of endorsement by OMERACT Community	○
19	Implement communication and dissemination plan	○

Assembly of working group and protocol development

1. Assemble working group

Following the same guidelines as used in the OMERACT Core Domain Set selection process, an OMERACT Working Group consists of a wide group of participants (at least two Patient Research Partners (PRP) must be named and active) along with a smaller number of individuals who form the Steering Committee. There are at least 3 co-chairs on the Steering Committee representing three different continents. In addition to the two PRP in the wider Working Group, one PRP must be a member of this Steering Committee. The Steering Committee is rounded out with a Fellow and at least two other members with expertise in the topic of the Working Group. Chapter 2 of this Handbook provides more details on establishing a Working Group.

The Working Group must involve PRP who are actively engaged in the project (e.g. participate in regular conference calls, provide expertise with lived experience of the disorder). PRPs can provide insight to various steps within the OMERACT instrument selection process, especially in the early stages of the process where PRP input on domain match and feasibility are critical. Chapter 3 of this Handbook, ‘Patient Partners and OMERACT’ provides further information on engaging patient research partners in a Working Group.

OMERACT has established a philosophy around the communication and engagement of members entitled the ‘Spirit of OMERACT’. This is outlined in detail in Chapter 1 of this Handbook. Working Groups are expected to foster the Spirit of OMERACT (e.g., collaboration, consensus) in all their work.

1	Assemble working group	○

2. Decide on methods protocol for Core Outcome Instrument Set selection

OMERACT has established an approach to instrument selection reflective of a long history in evidence-based decision making around our instruments (Boers 1998). The original Filter described the type of evidence we need, and has been in place since the inception of OMERACT. As more literature became available and methods advanced in OMERACT, measurement sciences and international groups in the field of core set development or measurement, OMERACT continually updated its processes and standards many times in conjunction with other international stakeholder groups (Beaton et al., 2020, Higgins et al., 2020, Prinsen et al., 2018, Mokkink et al., 2018). The result is the OMERACT Way (Figure 5.3), which describes an approach to see if a given instrument has “passed the OMERACT Filter” of ‘Truth’, ‘Discrimination’ and ‘Feasibility’. The step-by-step approach is outlined in the OMERACT Master Checklist for Instrument Selection as shown above. We have also developed an accompanying OMERACT Instrument Selection Workbook which outlines all the steps in the process and allows groups to record the results of their work. This prepared protocol is available for all groups to use. If a group wants to follow our protocol, then the workbook facilitates progress through the various steps. Our preference is that working groups follow the workbook approach that has been developed and approved by the OMERACT executive and community.

The OMERACT Way uses the following approach:

- Check for domain match and feasibility using methods described
- Seek agreement of the working group that this is an instrument that matches the target domain & is feasible
- Conduct a literature search using search terms available in the Appendix and modified for your need
- Select articles using screening and selection questions provided and create a PRISMA diagram
- Extract location of the evidence in a Summary of Measurement Properties (SOMP) table
- Conduct a good methods check (quality appraisal)
- Extract data on description of studies and results in our reporting summary tables templates

- Assess adequacy of the results of the measurement property evaluations
- Synthesize results for each measurement property (creating a profile across measurement properties)
- Complete a Summary of Measurement Properties (SOMP) Table to track the evidence
- Apply the OMERACT Algorithm to determine the recommended level of endorsement
- Present the evidence base to the OMERACT community for ratification

If a group decides to follow these methods, they simply tick the box in the instrument selection workbook acknowledging this. However, if the group decides to deviate from the proposed methods, we need to know ahead of time and the full Technical Advisory Group (TAG) will review these modifications. Working groups need to ensure any changes are well documented in the OMERACT Instrument Selection Workbook.

2	Decide on methods protocol for Core Outcome Instrument Set selection	○

3. Deliverable: Submit protocol using Instrument Selection Workbook to Technical Advisory Group [TAG]

The TAG is a group of methodologists, patient research partners, researchers, and statisticians who are part of OMERACT and volunteered to serve on this advisory group. As well as advising Working Groups on their progress through the OMERACT Filter or on the design of studies to fill gaps, they also help OMERACT to make sure we are using the best, most efficient methods for core set development. They are available to provide advice throughout the instrument selection process.

When the working group has acknowledged in the Instrument Selection Workbook that they will follow the methods outlined in the workbook, or have documented any changes they wish to make, this should be submitted to the OMERACT Secretariat (admin@omeract.org). The Secretariat will pass the protocol to the TAG.

3	**Deliverable:** Submit protocol using OMERACT Instrument Selection Workbook to Technical Advisory Group [TAG]	○

4. Review and approval of protocol by TAG

The TAG will review the protocol and provide comments on any proposed changes to the methods that the Working Group has requested. Once all TAG comments have been addressed by the Working Group and TAG has approved the final protocol, the group can move on to assessing the instrument under consideration.

4	Review and approval of final protocol by TAG	○

Review of evidence of instrument performance for existing or new instrument

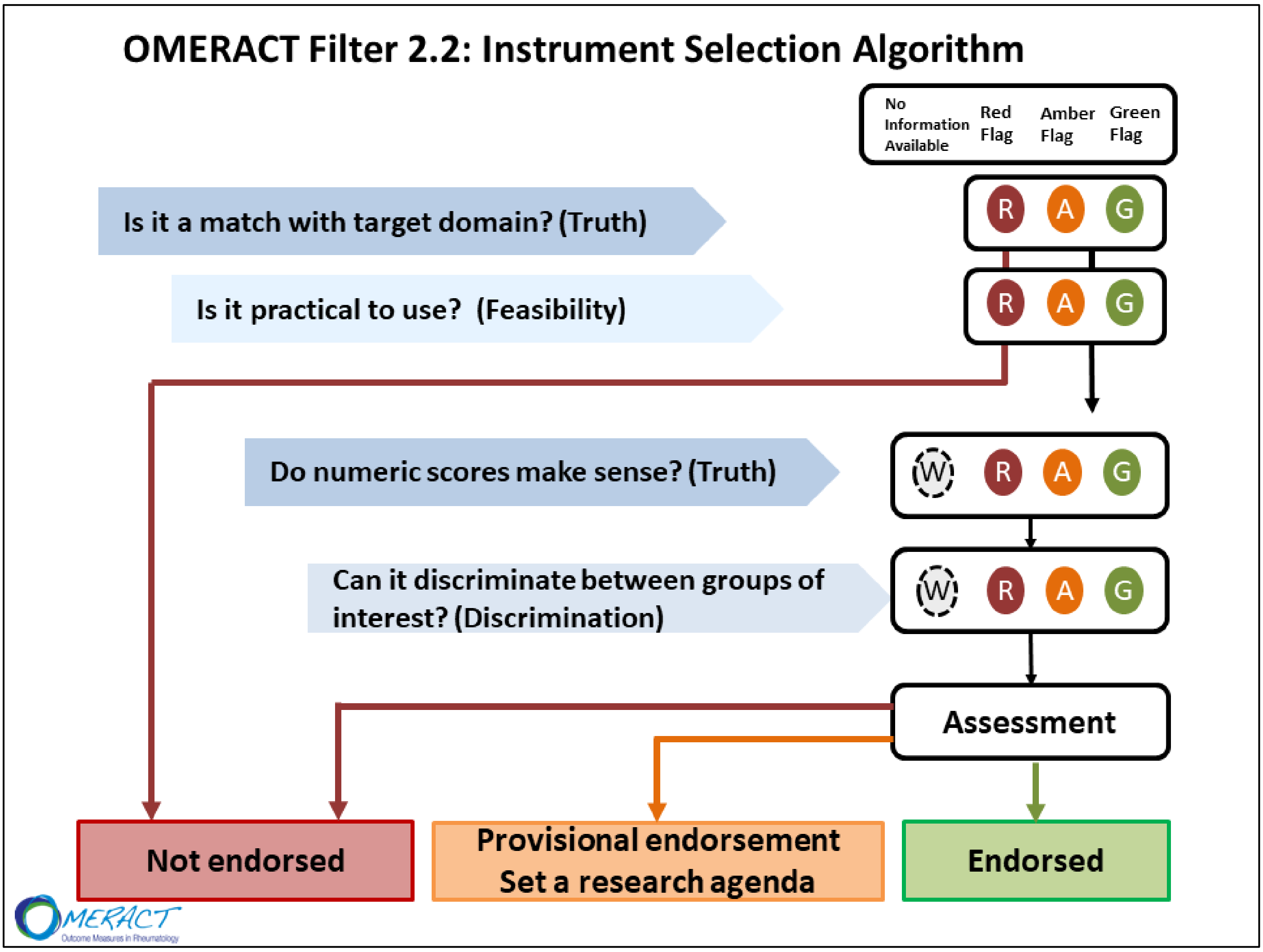

OMERACT Filter 2.2: Instrument Selection Algorithm
No Information Available
Red Flag
Amber Flag
Green Flag
Is it a match with target domain? (Truth)
Is it practical to use? (Feasibility)
Do numeric scores make sense? (Truth)
Can it discriminate between groups of interest? (Discrimination)
W
R
A
G
Assessment
Not endorsed
Provisional endorsement
Set a research agenda
Endorsed
OMERACT
Outcome Measures in Rheumatology

Figure 5.4 The OMERACT Filter 2.2 Instrument Selection Algorithm (OFISA) highlighting four questions to be answered for each instrument to move through the OMERACT Filter of Truth, Feasibility, and Discrimination.

Having received approval on the protocol, groups now move on to gather the information that is needed to answer the four questions in the OMERACT Filter 2.2 Instrument Selection Algorithm, also known by its acronym "OFISA", in Figure 5.4.

Each working group will have an OMERACT Senior Methodologist support person to help liaise with the TAG and is encouraged to communicate regularly with this person for advice and questions. The OMERACT methodologist supporting TAG, or the chair of TAG often attend working group meetings to offer ongoing support. In 2020, we initiated an open meeting called "Instrument Town Halls" in which anyone working on instrument selection can come onto this monthly call and work with other instrument selection working groups to address challenges or ask questions about the process. Periodically, we have working groups present an update of their progress at these meetings.

Part A: Domain match and Feasibility assessment

5. Obtain Working Group and others assessment of match with the target domain

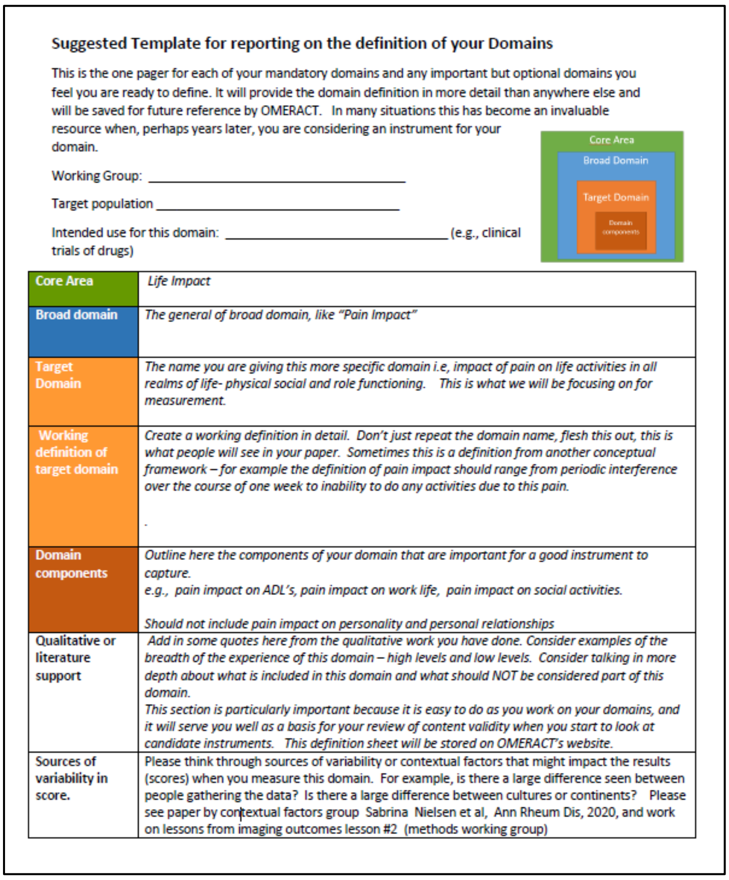
Is it a match with target domain? (Truth)

We begin the assessment of the instrument with the "*Truth*" pillar of the Filter. This addresses whether the instrument appears to be a good match for the target domain, and whether the instrument has the right content for the experience of that domain in the intended target population and study situation.

Essential to this assessment is the definition work done in the domain selection phase. Reviewing the domain definition template (see Figure 5.5) from the broad concept to the specific and focused target domain and its elemental components is important as an initial step to ensure that there is a match of the instrument with the definition previously established. To help you work through this material, we have compiled key references from the literature and have used them to develop sample survey questions (see Instrument Selection workbook). The working group should ask key stakeholders, including patient partners, about the domain match.

Key sources of information on evaluating domain match and feasibility:

- Auger C. Making sense of pragmatic criteria for the selection of geriatric rehabilitation measures. Ach Geronto and Geriatrics 2006:43;65-83.
- Feinstein AR. The theory and evaluation of sensibility. In Feinstein AR Clinimetrics. Westford MA: Murray Printing Co. 1987:141-166.
- Pakulis PJ. Evaluation physical function in an adolescent bone tumor population, Pediatr Blood Cancer 2005;45:635-643.
- Rowe BH., Oxman AD. An assessment of the sensibility of a quality-of-life instrument. Am J Emerg Med 1992;11(4);374-380.
- Smith M.L. Quality enhancement groups: A qualitative research method for survey instrument development. J Health Behav & Pub Health 2011:1(1);15-22.
- Terwee CB. Qualitative attributes of measurement properties of physical activity questionnaires: a checklist. Sports Med 2010;40(7):525-537.
- Terwee, C.B., Prinsen, C.A.C., Chiarotto, A. et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018; 27, 1159–1170



Suggested Template for reporting on the definition of your Domains

This is the one pager for each of your mandatory domains and any important but optional domains you feel you are ready to define. It will provide the domain definition in more detail than anywhere else and will be saved for future reference by OMERACT. In many situations this has become an invaluable resource when, perhaps years later, you are considering an instrument for your domain.

Working Group: ________________________________

Target population ______________________________

Intended use for this domain: ____________________________(e.g., clinical trials of drugs)

Core Area	*Life Impact*
Broad domain	*The general of broad domain, like "Pain Impact"*
Target Domain	*The name you are giving this more specific domain i.e, impact of pain on life activities in all realms of life- physical social and role functioning. This is what we will be focusing on for measurement.*
Working definition of target domain	*Create a working definition in detail. Don't just repeat the domain name, flesh this out, this is what people will see in your paper. Sometimes this is a definition from another conceptual framework – for example the definition of pain impact should range from periodic interference over the course of one week to inability to do any activities due to this pain.* .
Domain components	*Outline here the components of your domain that are important for a good instrument to capture.* *e.g., pain impact on ADL's, pain impact on work life, pain impact on social activities.* *Should not include pain impact on personality and personal relationships*
Qualitative or literature support	*Add in some quotes here from the qualitative work you have done. Consider examples of the breadth of the experience of this domain – high levels and low levels. Consider talking in more depth about what is included in this domain and what should NOT be considered part of this domain.* *This section is particularly important because it is easy to do as you work on your domains, and it will serve you well as a basis for your review of content validity when you start to look at candidate instruments. This definition sheet will be stored on OMERACT's website.*
Sources of variability in score.	Please think through sources of variability or contextual factors that might impact the results (scores) when you measure this domain. For example, is there a large difference seen between people gathering the data? Is there a large difference between cultures or continents? Please see paper by contextual factors group Sabrina Nielsen et al, Ann Rheum Dis, 2020, and work on lessons from imaging outcomes lesson #2 (methods working group)

Figure 5.5 Template for reporting detailed domain definition (*see OMERACT Handbook Chapter 4 for further details)*

Careful consideration is then given to the domain of the instrument and its global aim of the instrument, as well as the breadth and depth of the elemental components of the instrument; for example, item content in a PRO or what is visible in a specific imaging technique for "inflammation". This appraisal of the match with the target domain covers what is sometimes called *Face and Content Validity*. The assessment should include all perspectives: the patient, clinician and researcher perspective. If the instrument under consideration has different versions or different ways of scoring (for example, individual subscales versus the whole scale), the working group should clearly identify which version they are assessing.

At this stage it is also important to consider the sources of variability. Some of them are things we will talk about later in this chapter when discussing testing of their impact on scores, but some sources of variability should have been integrated into the domain definition itself, and are things that can be assessed at this stage of looking at the instrument. For example, when doing activities of daily living, one can think about doing them with and without the use of assistive devices. That is a source of variability and should be something your group is clear about in the levels of your domain definition. If you want to allow people to use assistive devices, you do not want to choose a questionnaire that forces people to respond without the use of an assistive device. Other examples are: do you allow pain assessment before or after a pain medication is taken? Do you allow people to assist someone in work activities when assessing a worker's productivity? Time of day can also be important if you are trying to measure morning pain or stiffness; you might want that to be measured in the morning if patients tell you that morning pain is the most important to them. You would reject a questionnaire that gathers data on night pain or even average pain rather than pain in the morning. These are all sources of variability that are identified and hopefully decided upon at the time of the creation of a detailed description of the definition of the target domain (further discussed in Chapter 4, section 6.3). These sources of variability are then carried forward to the instrument selection phase and checked on here under the candidate instrument's "match with target domain". Other sources of variability cannot be addressed by being more focussed in the definition. Things like the fact that two raters will be doing the assessments, or data will be gathered on two different imaging machines. These are likely sources of variability that will need to be tested in the section below called "inter-method reliability".

Traffic light scoring

Throughout the instrument selection process, "traffic light" scoring will be offered.

Green always means "good to go"

Amber always means there is a concern, or caution, or weakness but it is good enough to go forward.

Red always means stop, do not continue.

White means there is no evidence available

Example surveys and checklists for Working Groups are available in the instrument selection workbook and groups are encouraged to get multiple inputs – particularly from respondents about the adequacy of the content from the perspectives recommended by COSMIN: comprehensiveness, comprehensibility, and relevance of the content (Terwee et al., 2018).

We also encourage groups to examine some data of their own or from some publications to look at the distribution of responses, patterns of missing items, or floor and ceiling effects – all indicators of potential problems of the fit of the content with the population of interest. For imaging biomarkers, do the techniques (and proposed scoring) capture the intended pathophysiologic feature?

The result of the appraisal of domain match is then scored and recorded in the SOMP in the 'Domain Match' column using the traffic light formula of Green, Amber, or Red. The text box to the side provides the meaning of the traffic light scoring whenever it is used in the instrument selection process.

Example of Evaluation of Content Validity:
At the Patient Perspective Workshop at OMERACT 6, the concept of fatigue was identified by patients as an important outcome which was not included in the current RA core set. Further qualitative and quantitative studies explored the nature of fatigue as described by patients. Existing fatigue scales were found to omit many aspects of fatigue as reported by patients, and to include questions patients felt were unrelated to their fatigue experience. Therefore, a new fatigue scale, the Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional

Questionnaire (BRAF MDQ), was developed from items identified at interviews and focus groups with patients, followed by cognitive interviewing. Through this exercise, items and their wording were developed to cover a range of fatigue severity and impact. “Collaboration with patients enabled development of draft RA fatigue PROMs grounded in the patient data, strengthening face and content validity and ensuring comprehension.”

Kirwan J, Hewlett S. Patient Perspective: Reasons and Methods for Measuring Fatigue in RA. J Rheumatol 2007; 34: 1171–3.

Nicklin J, Cramp F, Kirwan J, Urban M, Hewlett S. Collaboration with patients in the design of patient-reported outcome measures: capturing the experience of fatigue in rheumatoid arthritis. Arthritis Care Res 2010; 62: 1552–8

5	Obtain Working Group and others assessment of match with the target domain	○

6. Obtain Working Group and others assessment of feasibility

Is it practical to use? (Feasibility)

The next step is the assessment of *Feasibility*. Feasibility includes those very practical considerations about cost, burden, access to the instrument in the necessary language(s), and mode of administration etc., that provides evidence to determine whether it is practical to use a given instrument. Input is needed from both the users of the instrument to comment on administration (researcher burden and cost issues), and from respondents to comment on burden and suitability of format and administration.

OMERACT requires an evaluation of the practicalities of using an instrument. Keeping the setting of the core set in mind, the cost, burden (patient, responder), equipment needs, sensitivity of content and overall ease of use are appraised (see optional appraisal forms you can choose to use if you wish in OMERACT Instrument Selection Workbook). Existing appraisal systems have been brought together to guide the types of questions to be asked of an instrument to assess its’ feasibility. Different perspectives should be included in this evaluation. As well as the clinician or researcher perspective, we feel direct input from patients is integral to the OMERACT Filter 2.2 process. Each offers an important contribution to the perspective of whether the tool is feasible (reasonable equipment needs, reasonable costs, training needs are feasible, comprehensive, easy enough to use). The decision makers will be balancing feasibility of the assessment process with having enough and the right content (items) to capture the full spectrum of the domain. Feasibility covers a broad array of factors and for more reading on this please refer to Tang et al., 2012; and Auger et al., 2006 review.

Examples of Feasibility Assessments:

Jensen 1986 et al. reported that NRS was “extremely quick and easy to administer”; only 5.3% of 75 patients made a mistake in the use of the scale.

Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: A comparison of six methods. Pain 1986; 27: 117-126.

Example: Tang et al, 2012 ran a survey to ask patients about their perceptions of a set of worker productivity measures. The tools were completed by the respondents and then they were asked about their preference of tool, and for each tool how difficult it was to reply, whether the items made sense, etc., and covered all important aspects, and whether the time to complete the tool was suitable.

Tang K, Beaton DE, Lacaille D, Gignac MA, Bombardier C. Sensibility of five at-work productivity instruments was endorsed by patients with osteoarthritis or rheumatoid arthritis. Journal of Clinical Epidemiology. 2013; 66(5): 546-56.

Example: In imaging, scan time must be just long enough to acquire sufficient information, but short enough to minimize radiation exposure, patient discomfort from immobilization, etc.

The result of the appraisal is then scored and recorded in the SOMP in the 'Feasibility' column using the traffic light system of Green, Amber, or Red.

6	Obtain Working Group and others assessment of feasibility	○

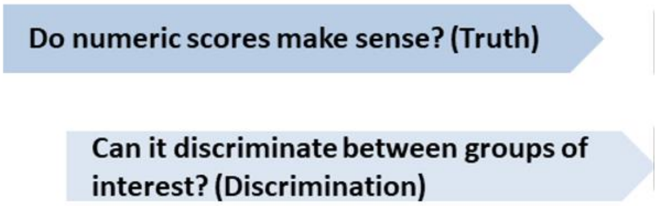
7. Obtain Working Group decision based on results of domain match & feasibility: Is the instrument a match with the domain AND feasible?

Decision point: Does the Working Group agree that this instrument has passed these first two questions?

We are now at an important decision point in the OMERACT instrument selection process. This decision point is a unique feature to the OMERACT process. If an instrument is not a good match for the target domain or is not feasible to use in the intended setting, it can be set aside by the Working Group. Ongoing attention should focus on only those instruments that have *passed* these two questions with a GREEN or AMBER rating. Many groups have found that a quick check of these first two steps eliminated several instruments that are covering the wrong content for the intended application, or are considered too long, expensive, and/or complex to use. It is best to set them aside and continue only with those that have content/concept match and are feasible to use in the intended application.

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)	○

Part B: Review of evidence of performance of an instrument across key measurement properties



Do numeric scores make sense? (Truth)

Can it discriminate between groups of interest? (Discrimination)

Once an instrument shows promise as matching the target domain and being feasible to use in the target context of use, the next step in the process is to conduct a systematic synthesis of the evidence to determine the measurement properties of the instrument. Evidence supporting the measurement properties of a given instrument is based on its ability to provide truthful and discriminative information as part of a core outcome measurement set. Parallel reviews are done for each property, and then evidence is synthesized across all properties to determine if there is sufficient evidence to support inclusion of the candidate instrument in a core outcome set (Boers , Kirwan, Gossec, et al., 2014, Boers, Kirwan, Wells, et al., 2014, Beaton et al., 2019). Many principles of a systematic review using

Cochrane methodology (Ghogomu et al., 2014, Higgins et al., 2020) and best evidence synthesis processes (Slavin et al., 1995) are followed in this stage. These include:

1. Formulation of a clear research question agreed upon through consensus of stakeholders;
2. Outline of a comprehensive and explicit search strategy;
3. Clearly defined eligibility criteria which are consistently applied;
4. Clearly defined data extraction which are consistently applied;
5. Rigorous and transparent critical appraisal of methods utilized and their results;
6. Synthesis methods that are explicit and appropriate.

> *Is there enough good evidence to suggest that the scores from this instrument are trustworthy representations of the target domain (concept of interest) for use in a core outcome measurement set?*

The question being asked of the instrument is "is there enough good evidence to suggest that the scores from this instrument are trustworthy representations of the target domain (concept of interest)?" In other words, does this instrument pass the OMERACT Filter of Truth, Discrimination, and Feasibility?

We recognize that ideally this comes studies found in the peer reviewed literature; however, when there are gaps in the literature, working groups often conduct their own study to develop evidence to fill in the gaps. Whilst these studies may have not yet undergone peer review, the TAG provides a critical appraisal of the work to offer some independent review. At the conclusion of the process, working groups are expected to present their evidence to the TAG and subsequently to the OMERACT community. In 2020, the OMERACT executive recognized that instrument selection was being held back by needing to wait for biennial meetings. They asked the TAG to review the results of the review in the form of a completed summary of the evidence, a workbook and tables of results. The TAG would comment on whether the methods used were consistent with the OMERACT guidance in this Handbook, and whether the results and conclusions of the working group could be supported by the data presented. Once the TAG has verified that the methods justify the conclusions, the working group's recommendation would go forward to the OMERACT community for ratification (more detail is provided in section 18 of this chapter).

The following sections describe the next steps of the process checklist and focus on this synthesis of the literature and the key components of it.

8. Conduct literature search; create PRISMA diagram; place articles of measurement properties in Summary of Measurement Properties (SOMP)

Working Groups (having approved an instrument's domain match and feasibility) now move into gathering evidence of the instrument's ability to perform as an indicator of the target domain in the intended context to answer the last two questions in OFISA, Figure 5.4; Question 3: "Do the numeric scores make sense?" (addressing the pillar of "truth" and the measurement properties of construct validity and reliability across testing situations – between raters, machines, etc., as necessary) and the four measurement properties that together address "Question 4: Can it discriminate between groups of interest"?

The four measurement properties addressing the pillar of "discrimination" are:

1. Stability in situations of no change (test-retest reliability)

2. Detecting instrument score change in situations of real change (longitudinal construct validity or responsiveness)

3. Discriminating between groups with the type of change anticipated in a clinical trial setting (a specific form of longitudinal construct validity or responsiveness)

4. Defining established thresholds of meaning

See Table 1 below for definitions for these terms as ratified by the International Society of Quality of Life Research (ISOQOL)(Reeve et al., 2013).

What are we aiming for?

For each measurement property, we are looking for consistent evidence (i.e., 2 or more) from studies that used good methods (reducing risk of bias) that show at least adequate level of performance of the instrument, in the absence of prevailing evidence suggesting the instrument is not performing well on this measurement property.

Most of this evidence will come from existing evidence in the published literature. We highly recommend registering a protocol for your systematic review of the literature search in a public register such as PROSPERO, a free online registry for systematic reviews. PROSPERO is available at: https://www.crd.york.ac.uk/PROSPERO/
Groups may choose to register their protocol in order to document their intent in this field. Having a registered protocol may also help with subsequent publication. Many of the fields in the Instrument Selection Workbook are the same information you need to register your protocol with PROSPERO.

Table 1. Definitions adopted for the indicators of performance of instruments moving through the OMERACT Filter 2.2 Instrument Selection Algorithm. (*Based on Reeve et al 2013, Feinstein 1987, and foundational work of Tugwell and Bombardier (Bombardier and Tugwell, 1987; Tugwell and Bombardier, 1982)*)

Filter 2.2 Pillar	Questions related to that pillar, corresponding measurement properties, and its definition
Truth (a)	*Is it a match with the target domain*? Content validity – the extent to which the instrument includes the most relevant and important aspects of the domain in the context of the intended measurement situation. Face validity - degree to which the instrument, as a whole, appears to be a match with the target domain.
Feasibility	*Is it feasible/practical to use?* Burden (respondent and researcher), time, effort, translations, and cost of using this instrument in the intended setting (context of use).
Truth (b)	*Do the numeric scores make sense?* Construct validity – the degree to which the scores on the instrument relate to other measures (patient-report or clinical indicators) in a manner that is consistent with theoretically derived, *a priori* hypotheses concerning the domains that are being measured, or show a distinction in scores between groups of known difference. Reliability across approaches – inter-rater reliability, inter-machine reliability, inter-setting reliability. This is a measure of the reproducibility of the instrument; the ability to provide consistent scores across sources of potential variability such as raters, machines, settings. There may be situation where this is not needed, as there are no important sources of variability. Please note, reliability over time is assessed under discrimination (test-retest reliability).

Discrimination	*Can it discriminate between groups in the setting of interest (clinical trial setting)?* Test-retest reliability – A measure of the reproducibility of the instrument, that is the ability to provide consistent scores over time in a stable population. Responsiveness (Longitudinal construct validity) - The extent to which an instrument can detect changes in the domain of interest over time, when they have occurred. Discrimination in Clinical Trials – The degree to which the instruments are sensitive to the related change between the arms of a trial (i.e., a comparative effectiveness study or a placebo-controlled trial, or active comparison arm). Thresholds of meaning for proportional summaries – The degree to which one can assign an easily understood meaning to the scores from an instrument. Rates of achievement are compared between arms in clinical trials. This includes thresholds like a minimum important improvement, or a patient acceptable symptom state.

Pause in the process: detailed descriptions of measurement properties

Let's pause here to think about each of these measurement properties and examples of the types of evidence we would find in the literature.

Construct validity

The next major question in OFISA asks whether the numbers/scores we obtain from the instrument are in line with what we would expect from our knowledge of the outcome domain which it is intended to measure. This is summarized as "Do the numeric scores make sense?". In many fields this is referred to as 'validity', moving away from more traditional label of construct or criterion validity. Either way, it is generally assessed by comparisons with the way the instrument of interest measures things in comparison to other instruments measuring the same or very similar domains (where they would be expected to give similar results) or measuring unrelated domains (where they might be expected to give different results). Evidence is ideally gathered in the same types of patients in similar situations (i.e., same technicians, same machine). Two very different domains both related to the underlying pathophysiology of a disease or condition are likely to correlate to some extent, so a judgment has to be made about the acceptable strength or weakness of correlations between instruments measuring similar and different domains to have it serve as evidence that the scores are making sense.

Establishing good comparisons is challenging. The *a priori* theories or expected results are based on knowledge of the domains (e.g. a link between pain and function might reasonably be expected in some conditions), the interventions (e.g. anti-depression treatment would not be expected to change x-ray findings), and/or the clinical disease itself (there may be relationships expected, e.g. signs of inflammation when people have active disease, but not when they have quiescent disease) in the studies being used to provide evidence of construct validity. The hypotheses of these expected results (including hypothesized thresholds for lack of correlation) should be formulated *before* the comparison analysis is performed or evaluated. Once hypotheses are described, the data is collected, and the statistical tests done to see if the findings confirm or refute the hypothesis. We have designed descriptive tables for extracting data on this type of validity evidence that try to help you extract this information. We know there will be gaps as often the *a priori* hypotheses describing the expected results are not described in a paper. This is particularly true of older literature. This could well be a problem of how the authors reported the study rather than what they actually did. However, because it is not reported, it becomes a risk of bias. Some working groups have identified that a study did not have the *a priori* hypotheses described in the article, but they could look at other literature or have a discussion with their working group members to decide on what they might expect. In these situations, it should be documented on the summary reporting table that this was not an *a priori* hypothesis from the article, but a working group estimation of the expected effect. Care should be taken as post-hoc analyses are prone to bias.

Construct validity is not an absolute property. It is always sensitive to the testing situation and to the patient groups involved in the testing (context of use). Confidence in how well the numbers obtained from an instrument/index represent the target domain is built by repeatedly obtaining pieces of evidence indicating that the numeric scores make sense. Consistency and congruency with other indicators will build confidence. The Filter 2.2 does not distinguish between construct and criterion validity. Criterion validity assumes that the comparator instrument is a "gold standard". Often there is no gold standard, and in that case the comparison is construct validity. Rather than entering into debates of the presence of a "gold standard" we suggest considering them robust constructs. There is one notable exception and that is in the area of biomarkers where they are validated for use as a diagnostic or predictive marker. In the latter case, the result that is being predicted (whether a clinical outcome or a result on another instrument) becomes the criterion. Similarly, in imaging, if one type of imaging is being used to be a proxy for the gold standard technique like magnetic resonance imaging (MRI), then the MRI might be considered the established gold standard. In the OMERACT Filter 2.2 we consider construct and criterion validity in the same category, varying only by the certainty of the comparator. Examples of criterion validity in the literature can be appraised under construct validity using the same method. Both answer our central question of whether the numeric scores make sense.

If there is insufficient evidence for construct validity in the proposed setting, the working group can make note of it and later can return to decide if they wish to undertake a study to obtain the necessary evidence.

Examples of construct validity

Example: To assess the construct validity of the newly developed Hip and Groin Outcome Score (HAGOS), the instrument developers hypothesized *a priori* how they expected the HAGOS would correlate with other scales that measure similar constructs. They hypothesized that the correlation between the HAGOS subscales 'Function in daily living' and 'Sport and recreation function' and the SF-36 subscale 'Physical functioning' was at least 0.50, and higher than for the other HAGOS subscales. The correlation between the HAGOS subscales 'Pain' and 'Symptoms' and the SF-36 subscale 'Bodily pain' should be at least 0.50 and 0.40 respectively, and higher than for the other HAGOS subscales. In a study of 101 patients the results were in accordance with the hypotheses; "As hypothesised, the correlations between the HAGOS subscales ADL and Sport/Rec and the SF-36 subscale PF were at least 0.5, and higher than for the other HAGOS subscales (Pain, Symptoms, PA and QOL). The correlations between the HAGOS subscales Pain and Symptoms and the SF-36 subscale BP were at least 0.5 and 0.4, respectively, and as hypothesised, higher than for the HAGOS subscales PA and QOL, but not higher than for the HAGOS subscales ADL and Sport/Rec.

Thorborg K, Holmich P, Christensen R, Petersen J, Roos EM. The Copenhagen Hip and Groin Outcome Score(HAGOS): development and validation according to the COSMIN checklist. Br J Sports Med 2011; 45: 478-491.

Example: To assess the construct validity of the Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP), the instrument developers hypothesized *a priori* how they expected the ICOAP would correlate with other pain, symptom and quality of life scales, as well as how results would differ between males and females. In a study of 100 patients, the authors found their hypotheses were validated; “The 11-item measure was significantly correlated, and in the directions expected, with the WOMAC pain scale, the KOOS symptoms scale, and self-rated affect of hip/knee problems on quality of life with Spearman correlation coefficients ranging in magnitude from 0.60 (KOOS symptoms) to 0.81 (WOMAC pain scale). As predicted, scores were slightly higher in women than men…”

Hawker GA, Davis AM, French MR, Cibere J, Jordan JM, March L, et al. Development and preliminary psychometric testing of a new OA pain measure: an OARSI/OMERACT initiative. Osteoarthritis and Cartilage 2008; 16: 409-414.

Cross sectional reliability across important sources of variability (inter-rater, inter-machine etc.)

When studying imaging types of outcomes, we learned that there are features in an assessment of an outcome that might influence the validity of the results. We refer to these as "sources of variability" in the scores obtained. The Contextual Factors working group recently created developed an operational definition of contextual factors and these sources of variability would fall into the "Measurement affecting contextual factors" (Nielsen 2021). These might include the differences between imaging machines, or between technicians training levels when conducting an ultrasound. We therefore recommend that groups consider the sources of variability that might be influencing why they are getting a certain score and test the consistency of scores obtained across that source of variability. A discussion of 'consistency of scores' should alert attention to the issue of reliability. In the situations above of differences between imaging machines or between technicians, the measurement property of interest would be inter-machine, or inter-rater reliability, respectively. We are thinking about the impact on cross-sectional scores which differentiates it from the test-retest reliability (in which reliability over time is of interest) that we will discuss in the next paragraph. Working groups need only look at cross-sectional reliability across important sources of variability. In some situations, this is not applicable, but should be considered, nonetheless.

Example of inter-rater reliability

Foppen et al. (2016) were interested in improving the reliability and agreement of scoring damage due to haemophilic arthropathy on X-rays, particularly in images with abnormal findings. They conducted a study where different raters interpreted the same set of X-ray images and provided their score based on that interpretation. They repeated the ratings using the same raters on a different set of images but this time offering a consensus atlas. The consensus atlas shows and describes examples of different images and how each should be scored; its purpose is to improve agreement, or consensus, between different readers. In this study, the authors found that the inter-rater reliability (a measure of agreement between two readers of the same X-ray) improved by providing a consensus atlas to the readers. Specifically, they found their intraclass correlation coefficient improved from 0.88 to 0.94 when the consensus atlas was used to guide the interpretation of the images and hence the final scores.

Foppen W, van der Schaaf IC, Beek FJA et al. Scoring haemophilic arthropathy on X-rays: improving inter and intra-observer reliability and agreement using consensus atlas. Eur Radiol 2016;26:1963-1970.

Can it discriminate between groups of interest? 1. Test-retest reliability: Are its scores stable when there is no change?

The next main question in OFISA is “Can it discriminate between groups of interest?”. In considering stability, test-retest reliability is an important property in order to ascertain how much day-to-day variability there is in the instrument score in a situation when the person or target should not be changing. Thus, attributes for a good study of test-retest reliability include: identifying a situation where no change in the target construct (e.g. pain intensity) is expected, and that there are repeated measures of the target construct over time, holding other variables constant (observer, time of day, etc.) as much as possible. Stability can subsequently be tested in other situations, e.g. with multiple observers.

The statistics that should be reported for stability are the ones that describe agreement (getting the same numbers out of the scale at both time points). This is provided by an intra-class correlation coefficient (ICC) or a weighted Kappa coefficient (K_w). There are different forms of ICC described by Shrout and Fleiss (1979), and the one most suited for most studies of test-retest reliability is the ICC (2,1) which is an ICC that is designed to deal with paired data (same people, two measures over time). Simple correlations run the risk of missing systematic differences in scores as they look for trends and not specific agreement in the actual numeric score. For example, if there were a learning effect and people always got 5 points higher on the second testing, this would be picked up in a statistic of concordance but missed in a correlation. It is important to look at the raw data (scatter plot of time one and time two; or 2x2 table).

Reliability coefficients sometimes run the risk of being calculated and then forgotten. But they can have really important information. ICC from test-retest reliability can be used to calculate the day to day variability in score. This is called the Minimum Detectable Change (MDC). The formula is fairly straightforward:

The MDC (at alpha level of significance) = z(alpha) x standard error of measurement (SEM) × √2. In the absence of the SEM, the formula: SEM= standard deviation (SD) x $(1\text{-}r)^{1/2}$. MDC = z(alpha) x √2 SD x √ (1-r) = z(alpha) x SD √(2(1-r)). At a 95% level of confidence, this would be MDC95 = 1.96x SD √(2(1-r)).

The minimum detectable change is the upper boundary of change in persons who did not change. That means that anyone who has more than this amount of change is not likely coming from this distribution of change in stable persons, i.e., you are likely to be seeing a true change. MDC is often described as a boundary of day to day variability in scores. Anything less than the MDC would be indistinguishable from just day to day variability.

Examples of test-retest reliability

Example: Ferraz, 1990: High test retest reliability has been reported in literate and illiterate patients with RA, r= 0.96, 0.95 respectively before and after medical consultation.

Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990; 17: 1022–4.

Example: Childs, 2005: “The test-retest reliability on the NPRS among patients whose condition had remained stable after one week resulted in an intraclass correlation coefficient of 0.61 ((0.30-0.77)….The 95% confidence intervals thus corresponds to a minimum detectable change of 1.99 points (1.96x1.02))(Table 2). These results indicate that a 2-point change on the NPRS is necessary to exceed measurement error based on a 95% CI.” (pg. 1333).

Childs JD, Piva SR, Fritz JM. Responsiveness of the Numeric Pain Rating Scale in Patients with Low Back Pain. Spine. 2005; 30(11): 1331–4.

Can it discriminate between groups of interest? 2) Longitudinal construct validity: Detecting change in situations of change (sometimes called "responsiveness")

Similar to construct validity (described above), an important part of understanding discrimination is to make sure that the scale can measure change accurately. Several situations of change (i.e., hypotheses on the *construct* of change) are formulated; these make up a set of 'mini-theories of change', stating expected direction and magnitude of changes in different groups or expected direction and magnitude of correlations between change scores of different instruments. Studies that enact these situations are then sought or conducted to generate evidence. Ideally, several comparators of good quality are included to verify that the expected change has indeed occurred. The change score on the scale is then compared to the results on the other instruments, and the formulated hypotheses. Several situations should be tested. Repeated concordance will build confidence in the ability of the instrument to measure change.

The appropriate statistics used to evaluate longitudinal construct validity depend on the how each study was set up. Often, standardized estimates of change (such as the effect size or the standardized response mean) are used and compared to an *a priori* estimation of the amount of change and direction that should be expected in the testing situation. (For example, in general, people having a joint replacement would be expected to have a large positive change in function, as indicated by a large effect size; this would then be expected in a score of hip function). The use of these statistics has been an area of controversy, particularly in readers of the COSMIN literature; however, careful reading would clarify that they only reject these statistics in situations where there are no *a priori* hypotheses or constructs of change. They would be appropriate to judge longitudinal construct validity in the presence of hypothesized situations of change, or when other anchors are used to estimate that the change has occurred (i.e., a larger effect size was seen in persons who said they had improved a lot compared to the effect size in people who were about the same). Other approaches include: assessing the correlation between change in a known valid indicator of change (global indices of change, other similar measures), and/or measuring the area under the curve (AUC) of the Receiver Operator Characteristic (ROC) curve if the known, valid referent measure can be dichotomized. The TAG can advise on and will pay attention to the details of the methods used.

Examples of longitudinal construct validity (responsiveness)

Example: Instruments to measure worker productivity were compared to an overall self-rating of change in ability to do their job as gathered on a global index of change (how much has your ability to do your job changed over the last three months?). Change scores in the Work Activity Limitation Scale (WALS) over the same three-month period were then correlated with the global indicator of change. The authors also estimated the effect size for people who had improved (WALS SRM = 0.79) and deteriorated (WALS SRM = -0.5) as another example of the validity of the change scores. Areas under the curve using "improved" versus "not improved" as the criterion were calculated and were 0.71 for improvement and 0.76 for deterioration suggesting good longitudinal construct validity.

Beaton DE, Tang K, Gignac MA, Lacaille D, Badley EM, Anis AH, Bombardier C. Reliability, validity, and responsiveness of five at-work productivity measures in patients with rheumatoid arthritis or osteoarthritis. Arthritis Care Res. 2010 ;62(1): 28-37.

Example: For evaluating responsiveness of the HAGOS, a Global Perceived Effect (GPE) score, where the patients (n=87) rate their condition in one of seven categories at 4-month follow-up, was used. A 4-month follow-up was chosen since this was a reasonably long timeframe to expect clinical improvement to occur in patients with long-standing hip and/or groin pain. It was hypothesised that the change in scores of the six subscales of the HAGOS between the initial administration and the 4-month administration would correlate with the GPE score, and that the correlation was at least 0.4 for all subscales. Furthermore, standardised response mean (SRM) and effect size (ES) should be higher for patients who reported their condition to be better or much better, than patients reporting no change, only somewhat better, or worse on the GPE score. SRM and ES should also be lower for patients reporting worse or much worse than patients reporting no change or only somewhat better or worse on the GPE score. All results were in accordance with the hypotheses.

Thorborg K, Holmich P, Christensen R, Petersen J, Roos EM. The Copenhagen Hip and Groin Outcome Score(HAGOS): development and validation according to the COSMIN checklist. Br J Sports Med 2011; 45: 478-491.

Can it discriminate between groups of interest? 3)Clinical trial discrimination: sensitivity to change in the context of a RCT

To gather information on how well an instrument will be able to perform in a clinical trial setting, groups must look for evidence from clinical trials rather than cross-sectional studies (i.e., the sensitivity to detect differences between the change in treatment A versus treatment B (or control) arms in a trial will be evaluated). The literature should be reviewed for studies where the instrument has been in a clinical trial. Often these are difficult to find but there might be information from the instrument's performance in a non-randomized study or in different cohorts. Published protocols often provide lists of primary and secondary outcomes. Industry partners also know of instruments that are fielded in drug or device trials. Some of this evidence could be gathered together to estimate, perhaps at a lower level of confidence, the likely ability of the instrument's ability to perform in a clinical trial. We considered these to be "gold, silver and bronze" levels of confidence and describe some examples of what could be found.

Gold standard assessments involve the placement of the candidate instrument in a clinical trial setting where other established, validated instruments are also completed so they can tell us the impact of the intervention upon the domain of interest. This setting is then used to test whether the trial results could be measured using the candidate instrument. This approach takes full advantage of the clinical trial setting, the change relative to true comparison groups, and the knowledge of the true effect outside of the measured effect on the measurement instrument.

At the silver level of evidence, the data comes from a two-arm, non-randomized study, so outside the context of the clinical trial setting. Patients were therefore not randomized to the two groups. The cohorts could still provide us with some likely effect size that could be observed in a clinical trial by measuring relative change etc., however, we recognize the potential for bias due to differences in the two groups, so this is a slightly more vulnerable to bias in the estimate.

At a bronze level, data from a single arm cohort can be divided into subsets of people who have and who have not improved, or who have not had a treatment response. The subsets are then compared for their relative change. This could involve asking people in a cohort whether their treatment worked according to a global estimate and then dividing the cohort into those who responded or not and compare change distributions (treatment effect sizes) between the groups.

Analysis includes estimates of the effect size in the two groups; and a responder analyses – the relative proportion of responders according to any established thresholds of change (e.g. 'minimum important difference' – see below).

Examples of clinical trial discrimination

Example of Gold evidence: In the context of the COBRA trial comparing sulfasalazine monotherapy to step-down combination therapy in early RA, both the (within-group) sensitivity to change and the (between-group) discrimination in change was assessed for a wide range of instruments. Overall, indices such as the ACR20 and the DAS performed better than single measures but ranking of sensitivity to change did not equal the ranking of between-group discrimination.

Verhoeven AC, Boers M, van Der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis 2000; 59: 966-74.

Example of Gold evidence: In a multicenter RCT comparing leflunomide, methotrexate and placebo, disease-specific and generic instruments to measure of function and health-related quality of life were assessed in RA patients. The relative efficiency of the instruments to detect a treatment effect relative to the tender joint count was assessed separately in the methotrexate versus placebo and leflunomide versus placebo groups. In comparing leflunomide with placebo, the patient global assessment, HAQ disability index, and SF-36 bodily pain scale were most responsive to treatment group differences. Authors concluded that "Both disease-specific and generic measures of function and health-related quality of life detect improvements in RA patients."

Tugwell P, Wells G, Strand V, Maetzel A, Bombardier C, Crawford B, et al. Clinical improvement as reflected in measures of function and health-related quality of life following treatment with leflunomide compared with methotrexate in patients with rheumatoid arthritis: Sensitivity and relative efficiency to detect a treatment effect in a twelve-month, placebo-controlled trial Arthritis & Rheum 2000; 43: 506-14.

Example of Silver evidence: If there are two cohort studies both using the instrument, although patients are not randomized, the two cohorts could give some indication of the relative effect that might be found in a RCT. For example, cohorts with joint replacement and conservative treatment of pain might be compared and a larger effect size might be expected at 6 months from the joint replacement group.

Example of Silver evidence: An important category is the non-inferiority study comparing a new treatment with one previously shown to be superior to placebo. The effect size of the instrument response compared to placebo might be inferred indirectly.

Example of Bronze evidence: In a single cohort, participants who change or who do not change according to a specific criterion (e.g. pre-defined responder analysis, patient self-report of definite improvement) the instrument response in the two groups can be compared. For example, in studying Worker Productivity measures, people who said they were successfully sustaining their work status might be compared to those who were feeling at risk of losing their job.

Can it discriminate between groups of interest? 4) Thresholds of meaning for individuals defined? Minimum important difference, Patient acceptable state

The final requirement in the OMERACT Filter is gathering and synthesizing the evidence on the thresholds of meaning for the scores i.e., how we should interpret the numeric scores. At OMERACT we ask for this to be provided in the SOMP even though others do not consider it a measurement property per se; however, we feel it is important information to have in hand for the ultimate use of the instrument in clinical research. There are several ways that we look at interpreting scores; often we focus on the meaning of the change in score but equally there are issues with benchmarking target levels of pain as an outcome. We will divide our discussion into two parts. Over the last 20 years since OMERACT began being a major force in advancing the methods to interpret outcome measurement instrument scores, there has been a lot of controversy as to the best approaches for determining a minimal clinically important change/difference (MCID) in scores (Beaton et al., 2001, Beaton et al., 2002, Tubach et al., 2005, Tubach et al., 2006, Wells et al., 2001). OMERACT has traditionally used the term MCID and more recently has adopted the term minimal important difference (MID) to represent change that is important for example to patients but may not have been generated in the context of its clinical relevance for clinical decision making. The overall approaches and its importance as a threshold of meaning remains unchanged. Thankfully, differences in opinion are converging and current best methods are emerging. We will focus on the methods we recommend.

a) Thresholds of meaning in change

When interpreting a change score, it is always good to also understand the background noise, or the day to day variability in scores that is due to measurement error alone. An interpretable change score is one that exceeds both a threshold of importance/relevant change, and this boundary of measurement error. This boundary has been called the smallest detectable change/difference (SDC/SDD) or the minimal detectable change/difference (MDC/MDD) or Jacobson's reliability change index (Jacobson et al., 1999, Stratford et al., 1996, Wells et al., 2003). It is directly related to the confidence interval calculated for the Bland-Altman plots of error (Bland and Altman 1986). This value converts a test-retest reliability coefficient into a change score that represents this error and we are asking groups to calculate it at the time that they are reviewing evidence on test-retest reliability. It is important to remember that this is a boundary of error. It is not a signal for "important" or "meaningful" difference for an individual.

The determination of "important" or "meaningful" change is the next threshold of meaning in change. In any of this work there are several questions to ask of the study:

o Who determined that this change was "important" or "meaningful"? Was it the patient, a clinician, a research team? The best studies use the source that relate to the respondent on the outcome. For PRO's that would be the patient, for interpreting an imaging outcome that could be the clinician or researcher.

o Operationally, how did they identify people that had a meaningful or important change? This is a critical element of a study. Anchors are needed to classify people as having experienced this meaningful change from those who have not. How credible is this process? How were they sure the change was "important"? How close will it get us to the target of an MCID for example?

o How did they then take these two groups (those who had an important change and those who did not) and decide on the threshold of change that would best discriminate between these two groups with the highest accuracy in doing so? This is the analytic approach for getting to the actual numeric score that will represent the MID.

With these three thoughts in mind, some approaches to determining important change will be challenged. Some promote the use of more distribution-based methods, such as the change score associated with an effect size of 0.5 (remembering that an effect size = mean change/standard deviation (SD)(baseline). To calculate this MID one would only need to know the standard deviation (SD). Although very attractive, and sometimes conveniently close to an approach that includes an anchor (Beaton 2003, Norman 2003) distribution-based methods are now considered suboptimal approaches for achieving MID values as there are no links to the meaning to the patient. More groups and regulators are favouring anchor-based approaches, approaches that when applied appropriately address all the points raised in the three bullet points above (FDA 2018). OMERACT methods ask for an anchor-based approach and requests that groups use multiple anchors to triangulate estimation of MIDs. We will now focus our discussion on the anchor-based approaches.

A key feature in a good study of MID is that the change captured is “important”; that is, it is valued in some way. One important consideration when looking at studies that purport to be capturing an important change is seeing how credible this “importance” assignment was. In some studies, the methods might use a global indicator of change and classify people who were one step above “no change”, say “a little bit better”, as having an important improvement.

The key here is that in some way that change needs to be “important” and not just “perceptible”. Studies must show that they have addressed the importance of the change when they are determining the MID.

Another consideration is that of the clinical relevance of the changes as well. This may or may not be the same as a minimal important change or difference to an individual patient. Clinical meaningfulness refers to an amount of change that is thought to be relevant to patients and/or clinicians and that has relevance for clinical decision-making. This could be achieved by the same or different anchor, and different input into how to use that anchor in classifying people as having had a clinically relevant change or not. In this case clinicians, patients and researchers might all be involved decision making, rather than prioritization of the patient’s or any other stakeholder’s view alone. Some have reserved the acronym MCID (minimal clinically important difference) for this type of change score.

The third bullet in our list, how the actual MID is determined is next to think about. The first thing that we know is that important improvements are different than important deteriorations and they should be evaluated and determined separately. Beyond that there are several approaches that have been used in the determination of the MID value. Some, like the mean score of people who have said they have had a small but important change on that anchor, are well used but do have some concerns. The mean value by definition is going to have a sizeable false negative rate as all patients in the distribution have all said they had a small but important change, and we are saying that the mean (likely in the mid-range of that distribution) is the threshold for an important change. Half the distribution, lower than the mean, would be considered as not having an important change.

The FDA has offered the following considerations for selecting the best anchors measure(s) considerations that are completely congruent with the OMERACT Way. (FDA 2019, PFDD Guidance 3 Discussion Document: Select, Develop or Modify Fit-for-Purpose Clinical Outcome Assessments, Section VI, D.1)

•Selected anchors should be plainly understood in context, easier to interpret than the clinical outcome assessment (COA or instrument) itself, and sufficiently correlated to the targeted COA.

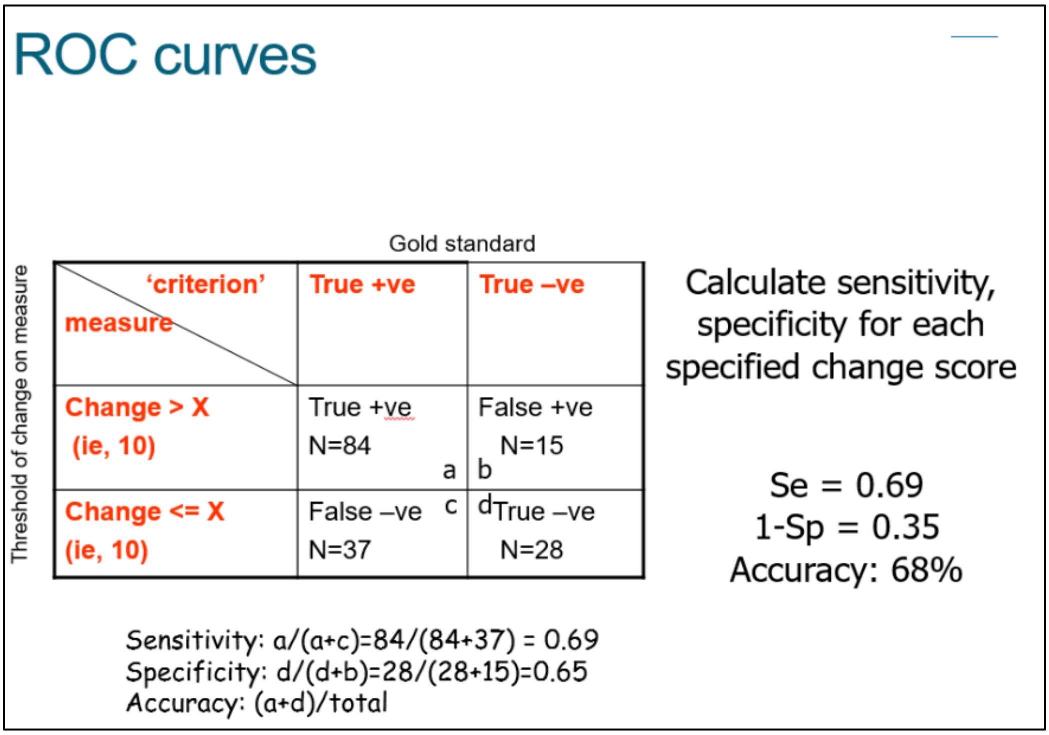
•Multiple anchors should be explored to provide an accumulation of evidence to help interpret meaningful within-patient score change which can also be a range.

•The following anchors are recommended to generate appropriate threshold(s) that represent a meaningful within-patient change in the target patient population:

- Static, current-state global impression of severity scale (e.g., patient global impression of severity or PGIS)
- Global impression of change scale (e.g., patient global impression of change or PGIC)
- Well-established clinical outcomes (if relevant)

•A static, current state global impression of severity scale is recommended at minimum, when appropriate, since these scales are less likely to be subject to recall error than global impression of change scales; they can also be used to assess change from baseline.

Other methods have been described for determining the actual change scores have been described in the literature (Youden indices, maximizing sensitivity and specificity). Given that MIDs are acting as classification of improved versus not improved groups, we recommend using approaches that have been advocated for diagnostic utility testing. In diagnostic testing the receiver operator characteristic (ROC) curve is used to plot the sensitivity and 1-specificity for each of several change scores on an instrument, where the “gold standard” is the anchor. The result is a ROC curve, showing the sensitivity and specificity of the various cut-points. The individual MID threshold is then chosen by one of several techniques such as the point with the highest diagnostic accuracy, the Youden index etc. This is done for each anchor being considered.



ROC curves

Gold standard

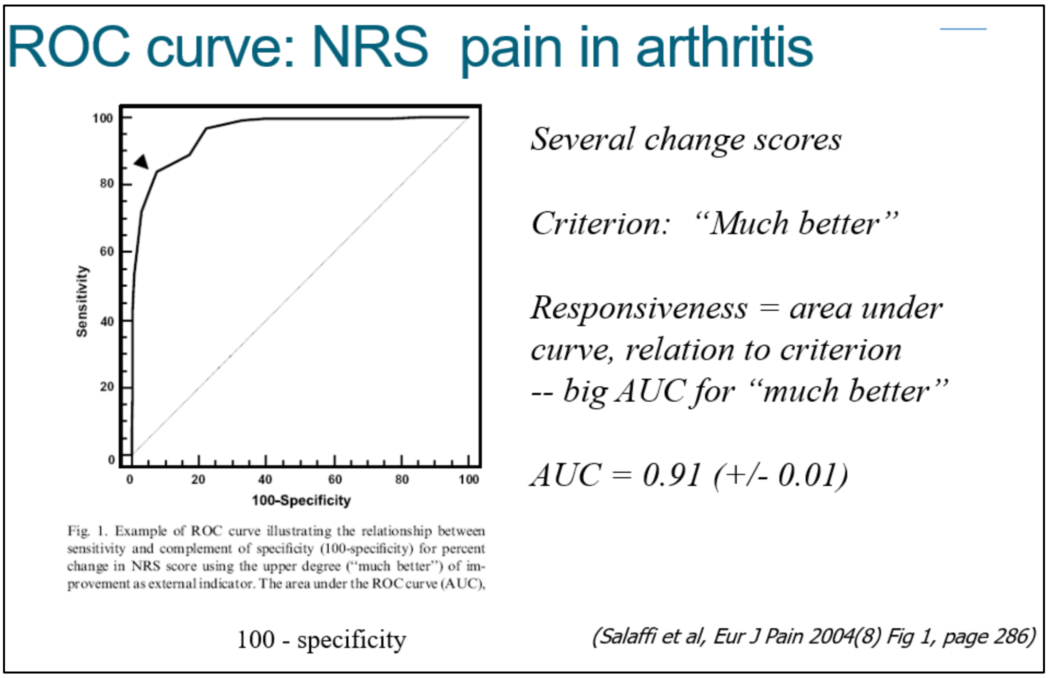
Threshold of change on measure

'criterion' measure	True +ve	True –ve
Change > X (ie, 10)	True +ve N=84 a	False +ve N=15 b
Change <= X (ie, 10)	False –ve N=37 c	d True –ve N=28

Calculate sensitivity, specificity for each specified change score

Se = 0.69
1-Sp = 0.35
Accuracy: 68%

Sensitivity: a/(a+c)=84/(84+37) = 0.69
Specificity: d/(d+b)=28/(28+15)=0.65
Accuracy: (a+d)/total



ROC curve: NRS pain in arthritis

Fig. 1. Example of ROC curve illustrating the relationship between sensitivity and complement of specificity (100-specificity) for percent change in NRS score using the upper degree ("much better") of improvement as external indicator. The area under the ROC curve (AUC),

100 - specificity

Several change scores

Criterion: "Much better"

Responsiveness = area under curve, relation to criterion -- big AUC for "much better"

AUC = 0.91 (+/- 0.01)

(Salaffi et al, Eur J Pain 2004(8) Fig 1, page 286)

More recently, advances have been made in the reporting of the MID values. The FDA (FDA 2018) asks for a cumulative distribution function overlaying the distributions for each level in the anchor. This permits one to see the distribution of individual change scores for each level of change and to see if each level is clearly representing a distinct distribution of change. These graphs provide transparency in the implication of the decision made in determining a MID as it will clearly show if the cut-off really was a clear, accurate cut-off or whether it overlapped tremendously with other thresholds.

Bookmarking Methods for defining either thresholds of change or state.

Recent efforts in bookmarking using nominal group processes to respond to patient scenarios and use rankings and ratings to define the split between patient scenarios with inconsequential disease manifestation versus mild (for example) and the difference instrument scores between these patient scenarios is examined and a cut point determined. This has most recently been used to define thresholds of symptom severity using PROMIS Measures in JIA. This method has also been applied to look at both severity thresholds as well as determining clinically meaningful clinical change in rheumatoid arthritis (Bingham CO, et al., 2021, Morgan et al., 2017). Of note in the later study, the amount of change that was reported as meaningful to patients and clinicians for fatigue and pain approximated or exceeded the amount of change associated with an anchor-based definition grounded on “a lot” of improvement or worsening (Bartlett et al., 2020).

Conclusion

MIDs are important thresholds because they will be used to classify people as being improved or not. MIDs are not simple. Different methods or different anchors will lead to different values. Recent guidance leans towards working with multiple anchors for this reason. MIDs have been found to vary for improvement and deterioration, and evidence should be sought for each in reviewing this literature. Finally, there is some concern that MIDs will vary by baseline distribution with some of this potentially being due to regression to the mean and others reflecting true differences in the amount of change that is important for people who are very ill versus just mildly ill. Many groups are working with MID by tertile of baseline distributions, or for various score ranges to overcome this.

Congruence across multiple indicators of MID (across different anchors, different patient severity, different directions of change) should help working groups build confidence in the estimation of this important threshold. Current guidance is to work with a range of MID values and conduct sensitivity analyses on whether the different threshold would alter response.

b) Thresholds looking at states

Another meaningful threshold is to identify benchmarks for scores at one point in time. For example, Tubach worked on the level of pain symptoms which would be considered “acceptable” to the patient and called this a “patient acceptable symptom state” (Tubach et al., 2006). Scores from the instrument were compared to an external anchor where patients have indicated if their current state would be acceptable to them if it were to continue unchanged. The distribution of persons who indicated it was acceptable was examined and a threshold determined. Perhaps the 70th or 90th percentile of the pain scores in those persons indicated an “acceptable state”. All the same principles we just discussed around MID will apply here, but in a cross-sectional manner. This analysis is reliant on the anchors chosen, and different anchors will produce different thresholds. So multiple anchors and triangulation of results should be done. The analysis could also make use of the ROC curve approach and the reporting using cumulative distribution functions to improve transparency.

Examples of thresholds of meaning

Example: Using the area under the curve analysis from responsiveness, Childs et al., chose the point to the highest sensitivity and specificity to reflect the best cut-off of change for pain. They concluded that a “2-point change on the NPRS (numerical pain rating scale) represents clinically meaningful change that exceeds the bounds of measurement error.” *Childs JD, Piva SR, Fritz JM. Responsiveness of the Numeric Pain Rating Scale in Patients with Low Back Pain. Spine. 2005; 30(11): 1331–4.*

Example: In two prospective studies to determine the minimal clinically important improvement (MCII) and patient acceptable symptom state (PASS), patients with knee OA and acute rotator cuff syndrome at follow up rated their current state as being acceptable if they were to stay in this state in an ongoing basis. “An anchoring method based on the patient’s response to therapy was used to determine the MCII and PASS.” “The minimal clinically important improvement was shown to be the change required to achieve the patient acceptable symptom state, whatever the baseline level of symptom, the outcome (pain or function), or type of condition (chronic or acute). This acceptable state for pain was higher for chronic (27.0 –36.4 across the baseline score) than acute (16.7–24.1) conditions.” *Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. Arthritis & Rheumatism 2006; 55: 526 –530.*

Example: Farrar et al used data from 10 placebo-controlled clinical trials of pregabalin in various chronic pain conditions to determine a minimum clinically important difference for the pain intensity (PI) NRS in chronic pain using ROC curves and the anchor of categories of ‘much’ and ‘very much’ improved in patient global impression of change (PGIC) categories. They found that “a reduction of approximately two points or a reduction of approximately 30% in the PI-NRS represented a clinically important difference”. *Farrar JT, Young JP, La Moreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measures on an 11-point numerical rating scale. Pain 2001; 94: 149-158.*

These pages have summarized some information about each measurement property that is contained in the OMERACT Filter 2.2, aiming for gathering evidence of an existing instrument’s ability to represent the target domain. Let’s return now to the process of getting and using this type of literature.

8.1 Searching for the evidence

Search strategies to find the evidence should be comprehensive, but focused. It has been shown that peer to peer mentoring of search strategy terms can improve the quality and comprehensiveness of searches (Sampson et al., 2009; McGowan et al., 2015). Feedback can be on terminology to reflect conditions, for example using arthrosis as well as arthritis, or improvements to the search strategy's use of Boolean logic or translation of search terms across databases. Working groups need input from a librarian or information specialist to design the best search strategy and for modifying the same for each of the databases. The search strategy provides the window into the literature. It needs to be done well.

For this purpose, we suggest a combination of three factors be combined in the search:

1. *Population.* Working Groups need to decide on the breadth of population they would like to consider (patient type, attributes, acuity, multiple diseases?). Note: some groups stick specifically with their target disorder or even a subgroup within that disorder such as early inflammatory arthritis while other groups include a broader array of disorders that are similar enough in their experience of this domain to offer relevant information.
2. *Instrument names, acronyms, or short forms* for the instrument should be set up to capture each time this instrument has been mentioned in the literature (either in title or abstract).
3. *Measurement properties*. We have provided search terms adapted from the COSMIN search strategies (Terwee et al., 2009) and offer them for use in several databases (see search strategies in the appendix of the Instrument Selection Workbook). The search will include terms beyond those measurement properties we have described above, but at this point, we suggest using this broader search in the case that additional measurement properties are hidden in that article. An article on factor analysis for example, might also include some construct validation.

Testing the search terms is highly recommended. Working Groups usually can find five articles that they know of on validity or reliability for their candidate instrument. Work with an information specialist to test and see if these are captured with the current search. If not, the information specialist can modify the search if needed to capture these key articles.

When each of *Population, Instrument, and Measurement properties* are defined and search strings created and tested, they are connected by Boolean ANDs to produce a much more focused intersection of these large search strings. The source of the evidence will be in the intersection. Often the addition of the population greatly reduces the yield and focuses on the most useful literature.

8.2 Screening and selection of articles

Quick screening of the articles can be conducted on titles and abstracts to make sure they are primary studies on measurement properties of the target instrument. Systematic or narrative reviews on the measurement properties of the instrument can be retained and the reference list can be checked to make sure all the relevant primary articles they included have been captured in your search. Screening questions are provided in the workbook.

At this phase, the full text of articles that have passed screening are obtained for a fuller review to verify they are primary studies about measurement performance of your instrument in a relevant population, and second to determine the measurement properties that will be addressed in that article if it is relevant. This phase is called selection. Two reviewers should conduct it and agreement is sought on both the inclusion of the article, and the properties it studies. It is quite easy to miss a measurement property during this stage. Most groups will be working with about 15-20 studies in the end, but each paper will usually provide evidence on multiple measurement properties. In the reviews done by MacDermid (2009); Schellingerhout (2012) such “measurement articles” were found to contain up to five measurement properties.

We suggest that, if possible, working groups use specialized software tools for screening and selection of articles. Word or Excel can be used but they are generally more time consuming. Two options for software programs that have templates for screening and selection are DistillerSR and Covidence; note that there is usually a fee associated for using these software programs.

8.3 Transparency in reporting your search results: PRISMA chart creation

Systematic reviews hinge on the understanding that they have looked widely and thoroughly for all the literature in a given area (population, intended purpose) and have carefully screened and selected the right articles. This means that all of the relevant literature will be influencing the synthesis of the findings and the final decision about the state of the evidence, in our case the decision about the quality of the instrument and whether it has passed the OMERACT Filter for application in the intended setting. Reporting the search and selection approach is critical to provide transparency in the process. Working groups should keep details of their search results as well as the screening and selection findings to allow them to fill in a PRISMA (Preferred Reporting Items for Meta-Analysis) flowchart (Moher et al., 2009). This can be done in Word or Excel; however, some working groups have used online software like DistillerSR and Covidence both of which can track selection and create a PRISMA chart for you. The end of the PRISMA flowchart (Figure 5.6) should have a level added below the final selected articles. This level presents a tally of the number of studies that presented evidence for each of the measurement properties. This number should match the number in the SOMP table.

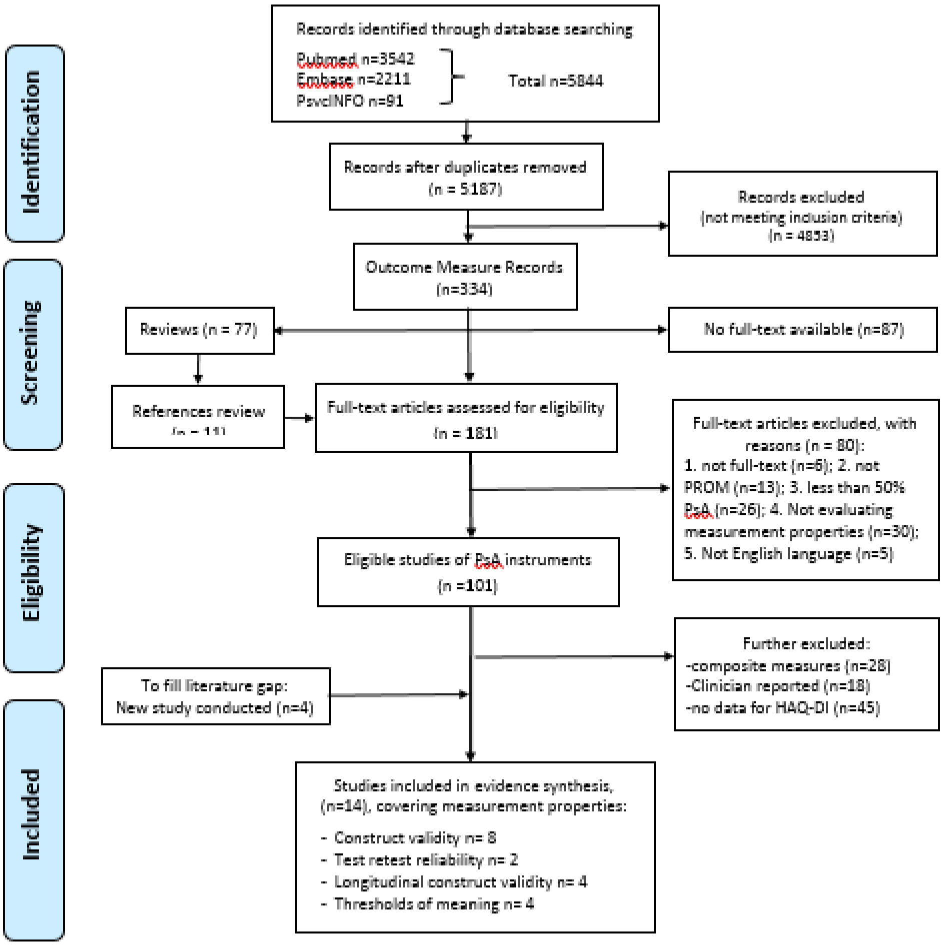

Identification
Screening
Eligibility
Included
Records identified through database searching
Pubmed n=3542
Embase n=2211
PsycINFO n=91
Total n=5844
Records after duplicates removed (n = 5187)
Records excluded (not meeting inclusion criteria) (n = 4853)
Outcome Measure Records (n=334)
Reviews (n = 77)
No full-text available (n=87)
References review (n = 11)
Full-text articles assessed for eligibility (n = 181)
Full-text articles excluded, with reasons (n = 80):
1. not full-text (n=6); 2. not PROM (n=13); 3. less than 50% PsA (n=26); 4. Not evaluating measurement properties (n=30); 5. Not English language (n=5)
Eligible studies of PsA instruments (n =101)
Further excluded:
-composite measures (n=28)
-Clinician reported (n=18)
-no data for HAQ-DI (n=45)
To fill literature gap: New study conducted (n=4)
Studies included in evidence synthesis, (n=14), covering measurement properties:
- Construct validity n= 8
- Test retest reliability n= 2
- Longitudinal construct validity n= 4
- Thresholds of meaning n= 4

Figure 5.6: Example PRISMA from PsA Working Group, HAQ-DI for physical function (*Leung et al. 2021*)

8.4 Tracking articles using the Summary of Measurement Properties Table

The articles identified as included in the PRISMA will then become the core of our review. We have created a table to help us track the studies, what they studied and what they found. We have called this the Summary of Measurement Properties Table or SOMP for short.

The Summary of Measurement Properties table will become the one-page summary of all your work! We will refer to this frequently as we move through the rest of the chapter and the instrument selection process. After filling in the top portion which includes a description of your intended application of the tool, the articles that were found in the literature review are placed in the rows, and the X's placed to identify which measurement properties were addressed in each article. A sum of the X's in the columns will identify the total number of articles available that could be giving us evidence for that measurement property. This should match the number on the bottom of the PRISMA Flowchart. If there is a zero or one, we will already know that more information will be required to meet our aim of having evidence from at least two good quality studies on each measurement property to tell us about its performance.

Later in the process, as the quality of the methods used in the study are checked (Good Methods Check), colour is added to each of the boxes to indicate if the reviewers determined this piece of evidence was conducted using good methods (GREEN OR AMBER) or not (RED). Empty boxes reflect WHITE, absence of information on that property from that study. X's are replaced with +, +/-, or – to indicate whether the findings of the study demonstrated adequate or better performance of the instrument (+), equivocal performance (+/-) or poor performance (less than adequate) (-). The evidence across those studies is reviewed and synthesized, and a Green, Amber, Red, or White (RAGW) rating is given to each measurement property (remembering domain match and feasibility were done and passed before the literature review and are shown here in the SOMP for completeness). The Working Group then decides what kind of endorsement they would like to present for a vote. Figure 5.7 is a fictitious example of a SOMP table. Each instrument has its own SOMP, specific to the intended context of use (patient population, type of intended study and comparators). In the rest of this section we will help you to build one.

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	Feasibility	Truth		Discrimination			
	Domain match		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.*							

Figure 5.7 Completed Summary of Measurement Properties (SOMP) table using a fictitious example

Now the working group will start to fill in their SOMP table by listing the studies they have found and placing an ‘X’ to show which measurement properties were assessed by each study.

Getting back to our process checklist, you will now be able to check off item 8:

8	Conduct literature search; create PRISMA diagram; place articles of measurement properties in Summary of Measurement Properties (SOMP)	○

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 <u>reported</u> methods, rather than <u>actual</u> 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 <u>not</u> be considered. In the COSMIN-OMERACT Good Methods Check, the reviewer then makes an overall decision about inclusion or exclusion of this evidence.			**Notes: (please keep notes about your ratings, and 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	○

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 MDC-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	○

11. Conduct synthesis across evidence available for each measurement property

All studies avoiding risk of bias in their design have now had their findings extracted and compared to the adequacy standards. The Working Group must now consider the synthesis of their information. OMERACT is using the best evidence synthesis approach blending Quality, Quantity, Consistency of findings, and Adequate (or better) Performance. This decision is guided by the work of others in best evidence synthesis groups (NQF 2013; Schellingerhout et al., 2012; Schmitt et al., 2015; Slavin, 1995). Best evidence synthesis looks for consistent evidence of good performance across multiple good quality studies of that property.

Quality has been determined at the level of quality appraisal as only those publications free of fatal flaws (GREEN, AMBER) are included in the synthesis. **Quantity, Consistency**, and **Adequacy** are now considered to complete the synthesis at this stage. For example, multiple high-quality studies could consistently show poor longitudinal construct validity of an instrument suggesting strong confidence against that measurement property for that instrument.

The literature gathered for each measurement property will be assigned a rating of GREEN (good evidence supporting this property, passes this element of the Filter), AMBER (some caution, or perhaps only one study on that property, but good enough to move forward) or RED (stop, evidence against this property or only poor-quality evidence) score. If there is no adequate quality evidence available on that property, it can be assigned a WHITE rating and await the creation of that evidence and future update of the rating.

Working Groups must gather all the evidence that they believe should be included in a synthesis for each of the six measurement properties required in the OMERACT Filter (construct validity, inter-method reliability, test-retest reliability, longitudinal construct validity, clinical trial discrimination, thresholds of meaning). Inter-method reliability (i.e., inter-rater, inter-machine) is new to Filter 2.2 to accommodate the lessons learned when integrating outcomes like imaging outcomes into OFISA. For these types of outcomes, the inter-rater reliability is a critical feature as there can be a lot of discordance between raters. In other situations, like a patient-reported outcome (PRO), sources of variability may not have been found and in that case the column will be marked NA (not applicable) and the related cell in the profile will be GREY and marked NA. This is not a weakness in the tool, just a measurement property that was not needed as a piece of evidence for that instrument.

The algorithm described in Figure 5.8 should be used as a guide for assigning the measurement property syntheses. A Green rating is assigned when there is consistent (at least two studies) evidence from studies with good enough quality supporting the instrument's performance in this measurement property. Note that the consistency of the evidence needs to be assessed across all the studies; it is not enough to find two studies with adequate evidence and decide not to continue reviewing the evidence - the entire body of evidence needs to be considered. A Red rating is assigned if there is an indication that this instrument is not performing well in this population and setting by demonstrating either inadequate findings in studies or if there are only studies deemed to not have good enough methods to provide credible evidence. White is assigned if there is no evidence available. Amber is assigned for all other situations.

Criteria for final rating								**Final rating for this measurement property**
Quality Of studies on measurement properties		*Quantity of good quality studies*		*Consistency across studies*		*Performance in this property*		
Good methods used	+	At least 2 pieces of evidence	+	Consistent findings	+	Adequate or better performance	→	GREEN
Good methods	+	At least 2	+	Consistent or Questionable	+	Inadequate performance	→	RED
Good methods	+	1 study only	…	NA	+	Inadequate performance		
Studies with fatal flaws	…	Not considered	…	Not considered	…	Not considered		
No evidence	…	0	…	NA	…	NA	→	WHITE
All other situations (Final rating not RED or GREEN or WHITE)							→	AMBER

Figure 5.8. Guide for synthesis ratings for each measurement property considering quality, quantity, consistency of findings across studies and adequacy of the performance on that measurement property.

The synthesis rating for each of the measurement properties are recorded on the Summary of Measurement Properties table in the "Synthesis Rating" row.

11	Conduct synthesis across evidence available for each measurement property	○

12. Identifying and managing gaps in the literature

12.1. Decide if any gaps exist in evidence of measurement properties

What if there are gaps (WHITE) or only methodologically flawed evidence (RED) in the evidence? If no other candidate instruments with better properties are available, new high-quality studies can be designed and performed (the TAG can help with design ideas) to fill gaps created by a lack of useable evidence (either absent, or only flawed evidence available).

12.2. If gaps found, draft protocol for new studies to fill gaps

When this synthesis is done, gaps may be identified (White in the RAGW ratings). If additional studies are needed to close those gaps, the working group should liaise with their OMERACT methods support person and the TAG will be engaged as mentors to assist with the design of the study that will avoid fatal flaws and provide the best evidence. The TAG has offered to review the protocol at this point to ensure you have considered the good methods checklist points in your design. An extra set of eyes always help. The working group is responsible for all good clinical practices in research including obtaining necessary research ethics board approvals prior to beginning the study.

These additional studies and their results will be moved through the same process as above as a new piece of evidence. If in the design phase the working group has made use of standards such as the COSMIN-OMERACT Good Methods Checklist (Beaton, 2019), COSMIN itself (Mokkink et al., 2010b; Terwee et al., 2012) and EMPRO (Valderas et al., 2008) of the SAC MOT guidance (Lohr, 2002), they will be more likely to have used good methods. This in conjunction with advice from the TAG will mean the study will be more likely to pass a subsequent review as having used “good enough methods” and avoid flaws related to a risk of bias.

While it is encouraged that these new studies are peer reviewed and published by the time of synthesis, we realize that some will not be published at the point when they are needed for OMERACT. The working group will provide a report on the design and findings of the study using the OMERACT templates for the reporting of each measurement property in a form very similar to a draft manuscript. Although risk of bias should be avoided with the deliberate design of the study to avoid it, a quick review of the results will be done by TAG members outside the working group to ensure the study would pass the COSMIN-OMERACT Good Methods Check. Results can then be entered into the synthesis step and the considered along with the published literature, but incorporating the rating given to the study in this review.

12.3. If no gaps exist, or if gaps cannot be filled, fill in SOMP Table with proposed level of endorsement of instrument

Working Groups now have a body of evidence that they feel is as complete as possible. Each measurement property has undergone the synthesis step described above and is represented by a GREEN, AMBER, or RED rating. Synthesis of this profile is then the final step in this process.

The algorithm described in Table 4 is used to determine the proposed level of endorsement: Endorsed, Provisionally endorsed, Not endorsed.

A GREEN in the synthesis rating row for every measurement property means a full endorsement of the instrument as having passed the OMERACT Filter 2.2.

A mixture of AMBER and GREEN ratings means provisionally passing the OMERACT Filter 2.2. When the recommendation is going to be AMBER (provisional), a statement of the work that needs to be done to bring it up to a full endorsement must also accompany it. AMBER is provisional not permanent. Working groups should commit to finding the remaining evidence and recognize that the completion of the evidence table could lead to a full endorsement OR to a decision that the instrument is not good.

Any WHITE ratings (a gap in the literature) or RED ratings (poor performance) found in the synthesis ratings across measurement properties means an instrument is lacking the supporting evidence and it would not be recommended for endorsement (do not endorse).

Table 4. OMERACT Algorithm to determine proposed level of endorsement

Full endorsement	All SOMP columns have a synthesis rating of GREEN. The instrument fulfils the requirements of OMERACT Filter 2.2 for inclusion in a core set.
Provisional Endorsement	There is a mixture of GREEN and AMBER synthesis ratings across the measurement properties. The instrument is endorsed for provisional inclusion in a core set until additional information is obtained. The working group sets a research agenda and continues to work on this instrument to see if it can become a fully endorsed instrument.
Not endorsed	Any of the columns have either RED or WHITE synthesis ratings. No available evidence, large gaps in evidence or flawed instrument performance suggest that this instrument does not yet have the evidence to support its use in a core set at this time.

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 overall rating of instrument	○

Initial submission to TAG: literature review findings & protocol for gaps

13. Deliverable: Submit the Instrument Selection Workbook to the TAG

By this time, the working group has amassed a large amount of information. The TAG now needs to review the reporting in the workbook, summary tables, PRISMA table, Good Methods Checklists, judgements for adequacy, and the SOMP with its synthesis ratings. The TAG is available along the way for any type of support but *must* be consulted at this point. Their job now is to conduct a methodological review of the work done to date.

Through a review of the workbook and supporting documents, and a discussion with the Working Group, the TAG looks at four things:

1. Evidence of the processes –search terms, PRISMA Flowchart, results, following through to the final SOMP and report.

2. Ratings and justification of the ratings at each stage.
3. Review any ongoing gaps that need to be filled and the protocol planned to fill that gap. They will be the peer review for any new work that the working group has to do to fill gaps.
4. Justification for the proposed final recommendation of the instrument and look at its consistency with OMERACT Filter 2.2 guidance.

The report to TAG must include the following:

1. The detailed definition of the target domain (from the domain selection process)

2. The completed Summary of Measurement Properties table synthesizing literature reviewed, scoring, profile for the instrument of synthesized findings for each property, and the proposed level of endorsement of the instrument.

3. The summary reporting tables of the studies and measurement properties.

4. The completed Instrument Selection Workbook [Appendix A] including the PRISMA flow chart and the detailed assessments of the COSMIN-OMERACT Good Methods check and assessments of the adequacy of the results (Excel or other format).

13	**Deliverable**: Submit the Instrument Selection Workbook to TAG	○

14. Receive final response from TAG

The TAG reviews protocols and completed workbooks for the methods that were used and the transparency of the reporting leading to the working group's conclusions. TAG looks for the pathway leading to the final conclusions and tries to ensure the OMERACT methods were followed. This can allow the OMERACT community to focus on the results of the review rather than being concerned about how it was done. The TAG does not endorse the results, they only state that the submitted documents show consistency with the methods of OMERACT and logically leads to the conclusions the working group has made. The working group is responsible for the results extracted from the papers and the conclusions they are drawing from them.

The TAG will provide the working group with comments on their submission and work with them to help answer any questions. This is an iterative process and the TAG is available to discuss with the working group either via email or teleconference calls.

14	Receive final response from TAG	○

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

When this synthesis is done, gaps may be identified (WHITE in the RAG ratings or AMBER when used to indicate only one available piece of evidence). If additional studies are needed to close those gaps, the TAG should be engaged as mentors to assist with the design of the study as described in section 12.b. Their input could guide working groups to avoid fatal methodological flaws and provide the best evidence.

These additional studies and their results will be moved through the same process as above. In the design phase, working groups would be wise to use the Good Methods Checklist, or other similar guidance on "best methods" such as COSMIN (Mokkink et al., 2010; Terwee et al., 2012), EMPRO (Valderas et al., 2008), the Scientific Advisory Committee of the Medical Outcomes Trust guidance (Lohr et al., 2002) or ISOQOL's guidance (Reeve et al., 2013) could be consulted in conjunction with advice from the TAG and used to ensure the inclusion of the key methodological features of a study needed to create and report a study that will be free of fatal flaws.

While it is encouraged that these studies are peer reviewed and published by the time of synthesis, we realize that some will not be published at the point when they are needed for OMERACT. The working group will provide a report on the design and findings of the study and will conduct data extraction. Although risk of bias should be avoided with the deliberate design of the study to avoid it, a review of the methods and results will be done by parties outside the working group to ensure the study would pass the Good Methods Checklist. This appraisal would be done by or with the TAG. Results can then be entered into the synthesis step and the considered along with the published literature.

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	○

Final submission to TAG for approval

16. Obtain agreement on final report

A final submitted report must be made that includes all the forms and reports submitted to the TAG in an iterative process, integrating any feedback until the TAG and the working group feel it is the final report.

16	Obtain agreement on final report	○

17. Set timeline for update of endorsed instrument

It is expected that instruments should be reviewed regularly to update the state of knowledge, with a maximum of 10 years. Working groups will be asked to set up a timeline for revisiting approved outcome instruments, and/or other potential emerging instruments. Working groups should consider means to store extracted review information in a manner to facilitate adding on more information at a future date.

17	Set timeline for next review of instrument	○

Ratification of level of endorsement by OMERACT Community and communication of results

18. Ratification of level of endorsement by the OMERACT Community

Consistent with the culture of OMERACT, we bring the body of evidence back to a wider community for final endorsement. For instruments, this will be done by sharing of key information through an online platform (this will include the domain definition worksheet, the PRISMA Flowchart, the SOMP, and the data extraction tables), and facilitating online discussion on a discussion board. Groups are encouraged to develop videos to help the community learn about their instruments and the results leading to their decisions. Following a 2-week discussion online, the working group will host an online meeting in order to summarize the results and address any key questions that arose on the discussion board.

Following the online meeting, attendees will be sent a voting survey asking them if they agree with the proposed recommendation that, based on the evidence presented, the [named instrument] be given a [Full endorsement, Provisional endorsement, Not endorsed at this time] for the domain of [x] in persons with [Condition]? Yes/No

If over 70% agree, the proposed level of endorsement will become the ratified level of endorsement. The results of this survey will be the record of endorsement.

18	Ratification of level of endorsement by OMERACT Community	○

19. Implement communication and dissemination plan

The final step in the OMERACT Filter 2.2 Instrument Selection process is to implement a communication and dissemination plan so that other stakeholders hear of the findings and Core Outcome Measurement Set. Planning for this as part of the process can help ensure time is allotted for ensuring stakeholders have access to the results. Consult knowledge translation (KT) expertise to think through creative ways to deliver your message effectively through publications, workshops, webinars, websites or other dissemination activities. OMERACT has a paper by Tunis et al, 2016, with valuable suggestions on KT approaches and recommends considering KT early in the process.

19	Implement communication and dissemination plan	○

Conclusion

The OMERACT Filter remains an important guidepost to the selection of instruments for core measurement sets in clinical trials. Truth, Discrimination, and Feasibility continue to be the pillars for decision making regarding the ability of an instrument to be endorsed for inclusion in a Core Outcome Measurement Set for clinical trials. In OMERACT Filter 2.2 we have updated the process to help OMERACT Working Groups deal with a large volume of information on measurement properties (some of good methodological quality and some not), advances in measurement sciences, and the need to synthesize findings across multiple studies. OMERACT Filter 2.2 builds on the experiences of many other international groups while retaining a clear link to the core OMERACT principles of evidence-based decision making, collaboration amongst key stakeholders, and consensus (Flurey et al., 2015).

Acknowledgements

The OMERACT Filter 2.2 instrument selection process reflects the evolution in the evidence-based decision-making processes that have been in place since the inception of OMERACT in 1992. We acknowledge the seminal work on the OMERACT Filter that took place at that and subsequent meetings (Boers, 1998). The template we have used for the systematic review of measurement properties has been influenced by the systematic review team at the Institute for Work & Health, Toronto, Canada and we thank Benjamin Amick, Kim Cullen and Emma Irvin for sharing their knowledge and templates with us. The Technical Advisory Group of OMERACT thanks Caroline Terwee for her thoughtful comments and contributions that was invaluable for our search strategy and our COSMIN-OMERACT good methods checklist. The review of adequacy criteria was extracted from a review of reviews of measurement property studies conducted by the Measurement Sciences group at the Institute for Work and Health as part of the Best Outcome Selection Studies and the OMERACT Handbook Team.

References

Auger C. Making sense of pragmatic criteria for the selection of geriatric rehabilitation measures. Ach Geronto and Geriatrics 2006:43;65-83.

Bartlett SJ, Gutierrez AK, Andersen KM, Bykerk VP, Curtis JR, Haque UJ, Orbai AM, Jones MR, Bingham CO 3rd. Identifying Minimal and Meaningful Change in PROMIS() for Rheumatoid Arthritis: Use of Multiple Methods and Perspectives. Arthritis Care Res (Hoboken). 2020 Nov 9. doi: 10.1002/acr.24501. Online ahead of print.PMID: 33166066

Beaton DE, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID working group. J Rheumatol. 2001;28:400–5

Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Current Opinion in Rheumatology 2002;14:109-14.

Beaton, D. Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. Medical Care. 2003; 41 (5) 593-6

Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, et al. Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. Journal of Clinical Epidemiology. 2011 May 1;64(5):487–96.

Beaton DE, Terwee CB, Singh JA, Hawker GA, Patrick DL, Burke LB, et al. A Call for Evidence-based Decision Making When Selecting Outcome Measurement Instruments for Summary of Findings Tables in Systematic Reviews: Results from an OMERACT Working Group. J Rheumatol. 2015; 42 (10) 1954-1961; DOI: 10.3899/jrheum.141446

Beaton DE, Maxwell LJ, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Instrument Selection Using the OMERACT Filter 2.1: The OMERACT Methodology. J Rheumatol. August 1 2019; 46 (8) 1028-1035; DOI: https://doi.org/10.3899/jrheum.181218

Beaton DE, Boers M, Tugwell P, Maxwell L. Chapter 36: Assessment of Health Outcomes. In: Firestein & Kelley's Textbook of Rheumatology. 11th Edition. Firestein GS, Budd RC, Gabriel SE, McInnes IB, O'Dell JR, Koretzky G (ed). 2020. Elsevier. ISBN-13: 978-0323639200

Bingham CO, Butanis AL, Orbai AM, Jones M, Ruffing V, Lyddiatt A, Schrandt MS, Bykerk VP, Cook KF, Bartlett SJ. Patients and clinicians define symptom levels and meaningful change for PROMIS pain interference and fatigue in RA using bookmarking. Rheumatology (Oxford), 2021; Jan 20:keab014. doi: 10.1093/rheumatology/keab014. Online ahead of print. PMID: 33471127

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement, Lancet, 1986; 307-10

Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol. 1998 Feb;25(2):198-9.

Boers M, Idzerda L, Kirwan JR, Beaton D, Escorpizo R, Boonen A, et al. J Rheumatol. Toward a Generalized Framework of Core Measurement Areas in Clinical Trials: A Position Paper for OMERACT 11 2014; 41 (5) 978-985; DOI: https://doi.org/10.3899/jrheum.131307

Boers M, Kirwan JR, Gossec L, Conaghan PG, D'Agostino MA, Bingham CO, et al. How to Choose Core Outcome Measurement Sets for Clinical Trials: OMERACT 11 Approves Filter 2.0. J Rheumatol. 2014, 41 (5) 1025-1030; DOI: 10.3899/jrheum.131314

Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. J Clin Epidemiol 2014;67:745-53.

Boers M, Beaton DE, Shea BJ, Maxwell LJ, Bartlett SJ, Bingham III CO, et al OMERACT Filter 2.1: Elaboration of the Conceptual Framework for Outcome Measurement in Health Intervention Studies. J Rheumatol. August 1 2019, 46 (8) 1021-1027; DOI: https://doi.org/10.3899/jrheum.181096

Bombardier C, Tugwell P. Methodological considerations in functional assessment. The Journal of rheumatology. Supplement. 1987 Aug;14 Suppl 15:6-10.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ. 2003;326(7379):41–4.

Devji, T, Carrasco-Labra A, Qasim A et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ (Clinical research ed). 2020;369:m1714

Duarte-García A, Leung YY, Coates LC, Beaton D, Christensen R, Craig ET, et al Endorsement of the 66/68 Joint Count for the Measurement of Musculoskeletal Disease Activity: OMERACT 2018 Psoriatic Arthritis Workshop Report. J Rheumatol. 2019 Aug;46(8):996-1005. doi: 10.3899/jrheum.181089.

Engel L, Beaton DE, Touma Z. Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation. Rheum Dis Clin North Am. 2018 May;44(2):177-188. doi: 10.1016/j.rdc.2018.01.011. Epub 2018 Feb 21. PMID: 29622290

FDA. Discussion Document for Patient-Focused Drug Development Public Workshop on Guidance 3: Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments, Workshop, October 2018. https://www.fda.gov/media/116277/download

Feinstein AR. The theory and evaluation of sensibility. In Feinstein AR Clinimetrics. Westford MA: Murray Printing Co. 1987:141-166.

Ghogomu E, Maxwell LJ, Buchbinder R, Rader T, Pardo Pardo J, Johnston R, Christensen R, Singh J, Wells GA, Tugwell P and The Editorial Board of the Cochrane Musculoskeletal Group. Updated Method Guidelines for Cochrane Musculoskeletal Group Systematic Reviews and Meta-Analyses. J Rheumatol 2014 Feb;41(2):194-205. doi: 10.3899/jrheum.121306

Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ [Internet]. 2011;343. Available from: https://www.bmj.com/content/343/bmj.d5928

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane, 2020. Available from www.training.cochrane.org/handbook.

Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: Description, application, alternatives. Jounral of Consulting and Clinical Psychology 1999;67(3):300-7.

Kirkham JJ, Boers M, Tugwell P, Clarke M, Williamson PR. Outcome measures in rheumatoid arthritis randomised trials over the last 50 years. Trials. 2013 Oct 9;14(1):324.

Kirkham JJ, Clarke M, Williamson PR. A methodological approach for assessing the uptake of core outcome sets using ClinicalTrials.gov: findings from a review of randomised controlled trials of rheumatoid arthritis. BMJ 2017; 357:j2262

Kirwan JR, Boers M, Hewlett, Beaton D, Bingham CO, Choy E, et al. Updating the OMERACT Filter: Core Areas as a Basis for Defining Core Outcome Sets. J Rheumatol. May 2014, 41 (5) 994-999; DOI: 10.3899/jrheum.131309

Leung YY, Orbai AM, Hojgaard P, Holland R, Mathew A, Goel N, Chau J, et al. HAQ-DI and the SF-36 Physical Functioning subscale provisionally endorsed as outcome measurement instruments of the physical function domain in psoriatic arthritis using OMERACT methodology. Seminars in Arthritis and Rheumatism. 2021 (in press)

Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979–92.

Lohr K. Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research. 2002;(11):193-205

MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, Goldsmith CH. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009 May;39(5):400-17. doi: 10.2519/jospt.2009.2930. PMID: 19521015.

Maxwell LJ, Beaton DE, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Core Domain Set Selection according to OMERACT Filter 2.1: The 'OMERACT Way'. J Rheumatol; Aug 2019, 46 (8) 1014-1020; DOI: 10.3899/jrheum.181097

McHorney CA., Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4(4):293–307.

McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016 Jul 1;75:40–6.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737-45. doi: 10.1016/j.jclinepi.2010.02.006. PMID: 20494804.

Mokkink, L.B., de Vet, H.C.W., Prinsen, C.A.C. et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018 (27) 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010;19:539-549. doi: 10.1007/s11136-010-9606-8.

Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology 2010;10:22. doi: 10.1186/1471-2288-10-22

Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, Knol DL, Bouter LM, de Vet HCW. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Quality of Life Research 2009;18:313-333. doi: 10.1007/s11136-009-9451-9

Mokkink et al. 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 https://doi.org/10.1186/s12874-020-01179-5

Morgan EM, Mara CA, Huang B, Barnett K, Carle AC, Farrell JE, Cook KF. Establishing clinical meaning and defining important differences for Patient-Reported Outcomes Measurement Information System (PROMIS()) measures in juvenile idiopathic arthritis using standard setting with patients, parents, and providers. Qual Life Res. 2017 Mar;26(3):565-586. doi: 10.1007/s11136-016-1468-2. Epub 2016 Dec 2.PMID: 27913986

Nielsen SM, Uggen Rasmussen M, Boers M, et al. Towards consensus in defining and handling contextual factors within rheumatology trials: an initial qualitative study from an OMERACT working group. Annals of the Rheumatic Diseases 2021;80:242-249

Norman GR, Sloan JA, Wyrwich KW. Interpreation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical Care 2003; 41, 582-92

Nunnally, J. C., & Durham, R. L. (1975). Validity, reliability, and special problems of measurement in evaluation research. In E. L. Struening, & M. Guttentag, (Eds.). Handbook of evaluation research. (Vol. 1). (pp. 289-352). London: Sage Publications.

Nunnally J, Bernstein JC (1994). Psychometric theory.3rd edn. New York: McGraw-Hill.

NQF 2013. Review and Update of Guidance for Evaluating Evidence and Measure Testing. Technical Report. National Quality Forum. 2013 https://www.qualityforum.org/Publications/2013/10/Review_and_Update_of_Guidance_for_Evaluating_Evidence_and_Measure_Testing_-_Technical_Report.aspx

Pakulis PJ. Evaluation physical function in an adolescent bone tumor population, Pediatr Blood Cancer 2005;45:635-643.

Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May;27(5):1147-1157. doi: 10.1007/s11136-018-1798-3. Epub 2018 Feb 12. PMID: 29435801; PMCID: PMC5891568.

Reeve, B.B., Wyrwich, K.W., Wu, A.W. et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 22, 1889–1905 (2013). https://doi.org/10.1007/s11136-012-0344-y

Revicki D, Hays RD, Cella D, et al. . Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9. 10.1016/j.jclinepi.2007.03.012

Roofeh D, Barratt S, Wells AU, Kawano-Dourado L, Tashkin D, Strand V, et al. Forced vital capacity endorsed by OMERACT Filter 2.1 as an outcome measurement instrument of lung physiology in systemic sclerosis associated interstitial lung disease: a systematic review. Seminars in Arthritis and Rheumatism. 2021 (in press)

Rowe BH., Oxman AD. An assessment of the sensibility of a quality-of-life instrument. Am J Emerg Med 1992;11(4);374-380.

Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009;62:944-952

Schellingerhout, J.M., Verhagen, A.P., Heymans, M.W. et al. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res 21, 659–670 (2012). https://doi.org/10.1007/s11136-011-9965-9

Schmitt J, Apfelbacher C, Spuls PI, Thomas KS, Simpson EL, Furue M, et al. The harmonizing outcome measures for eczema (home) roadmap: A methodological framework to develop core sets of outcome measurements in dermatology. J Invest Dermatol 2015;135:24-30.

Scientific Advisory Committee, Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002, 11:193–205. doi: 10.1023/A:1015291021312. [PubMed: 12074258]

Shrout PE and Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychol Bull 1979; 86(2):420-428.

Slavin RE. Best evidence synthesis: An intelligent alternative to meta-analysis. J Clin Epidemiol 1995;48:9-18.

Smith M.L. Quality enhancement groups: A qualitative research method for survey instrument development. J Health Behav & Pub Health 2011:1(1);15-22.

Stratford PW, Binkley J, Soloman P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland- Morris questionnaire. Phys Ther 1996;76(4):359-68

Tang K, Beaton DE, Lacaille D, Gignac MA, Bombardier C. Sensibility of five at-work productivity instruments was endorsed by patients with osteoarthritis or rheumatoid arthritis. Journal of Clinical Epidemiology. 2013; 66(5): 546-56.

Terwee CB. Qualitative attributes of measurement properties of physical activity questionnaires: a checklist. Sports Med 2010;40(7):525-537.

Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of Life Research 2012;21(4):651-7.

Terwee, C.B., Prinsen, C.A.C., Chiarotto, A. et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 27 2018; 1159–1170 https://doi.org/10.1007/s11136-018-1829-0

Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis 2005;64:34–37. doi: 10.1136/ard.2004.023028

Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. Arthritis Rheum 2006;55:526–30. https://doi.org/10.1002/art.22110

Tugwell P, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. J Rheumatol. 1982 Sep-Oct;9(5):758-62.

Tugwell P, Boers M. OMERACT Conference on Outcome Measures in Rheumatoid Arthritis Clinical Trials: Introduction. J Rheumatol 1993;20:528-30.

Tugwell, P., Boers, M., Brooks, P. et al. OMERACT: An international initiative to improve outcome measurement in rheumatology. Trials 8, 38 (2007). https://doi.org/10.1186/1745-6215-8-38

Tugwell P, Boers M, D'Agostino M, Beaton D, Boonen A, Bingham CO, et al. Updating the OMERACT Filter: Implications of Filter 2.0 to Select Outcome Instruments Through Assessment of "Truth": Content, Face, and Construct Validity. J Rheumatol. May 2014, 41 (5) 1000-1004; DOI: 10.3899/jrheum.131310

Tunis SR, Maxwell LJ, Graham ID, Shea BJ, Beaton DE, Bingham CO, et al. Engaging Stakeholders and Promoting Uptake of OMERACT Core Outcome Instrument Sets. J Rheumatol. 2017, 44 (10) 1551-1559; DOI: 10.3899/jrheum.161273

Valderas JM, Ferrer M, Mendívil J, Garin O, Rajmil L, Herdman M, Alonso J; Scientific Committee on "Patient-Reported Outcomes" of the IRYSS Network..Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health. 2008 Jul-Aug;11(4):700-8. doi: 10.1111/j.1524-4733.2007.00309.x. Epub 2008 Jan 8.

Wells GA, Beaton DE, Shea B, Boers M, Simon L, Strand V, et al. Minimal clinically important differeneces: Review of methods. J Rheumatol 2001;28(2):406-12

Wells GA, Boers M, Shea B, Anderson JJ, Felson D, Johnson K, et al. MCID/Low disease activity state workshop: Low disease activity state in rheumatoid arthritis. J Rheumatol 2003;30(5):1115-8.

Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of Variation and Bias in Studies of Diagnostic Accuracy. Ann Intern Med. 2004 Feb 3;140(3):189–202.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155: 529-36.

Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E., et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012;13-132.

Wyrwich KW, Norquist JM, Lenderking WR, Acaster S, the Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). Methods for interpreting change over time in patient-reported outcome measures. Quality of Life Research. 2013 Apr 1;22(3):475–83.

Frequently Asked Questions

How does the OMERACT Filter 2.2 Instrument Selection Algorithm (OFISA) relate to the development of new instruments?

The OMERACT Filter 2.2 Instrument Selection Algorithm is a process for finding and synthesizing findings of measurement properties to aid in instrument selection for already existing instruments. It is not a description of instrument development. Many other considerations go into instrument selection including item / attribute choices, framing of questions, and decisions on structure, scoring and scaling. This requires a different set of skills and methods. Many OMERACT Working Groups have chosen to develop a new instrument because no instrument was available, or the existing instruments could not meet OMERACT Filter 2.2 requirements and received RED or WHITE ratings. Instrument development is a long, labour intensive process. Development of a new instrument is therefore beyond the scope of the current chapter, and should only be undertaken if absolutely necessary, the measurement properties of the new tool will need to pass through the full Filter 2.2 requirements to verify that they have enough evidence gathered together to satisfy the Working Group and the OMERACT community that they have moved through the steps described above landing with a synthesis statement of GREEN or AMBER. As described above, the TAG would help the Working Group with the integration of unpublished work into OMERACT Filter 2.2 Instrument Selection evidence.

In sum, the same requirements need to be met for existing and new instruments developed for OMERACT. All need to be appraised for Truth, Discrimination, and Feasibility, but for new instruments, evidence does not need to be already published.

Appendices

The OMERACT Handbook group created workbooks with detailed search strategies and checklists to help Working Groups move through the steps outlined above. The workbooks facilitate gathering enough information to allow groups to register their review on public platforms for reviews such as PROSPERO (https://www.crd.york.ac.uk/PROSPERO/) . Groups are encouraged to do so to facilitate transparency and the publication of the results in the future.

We hope that the accompanying workbooks and appendices help with tracking the steps and organizing information for your own use in publications, and in presentations back to the OMERACT community.

List of Appendices

A. Instrument Selection Workbook for documenting process of gathering and synthesizing evidence.

CLICK HERE

B. The COSMIN-OMERACT Good Methods Checklist adapted for OMERACT Filter 2.2 Instrument Selection needs.

CLICK HERE

C. Data extraction templates for reporting each measurement property

CLICK HERE

D. Adequacy of results review

In development