OMERACT Instrument Selection

Topic: Sources of variability

This document provides readers with a guide to various resources on identifying and testing sources of variability using OMERACT Instrument Selection methodology.

A. Guidance to testing sources of variability

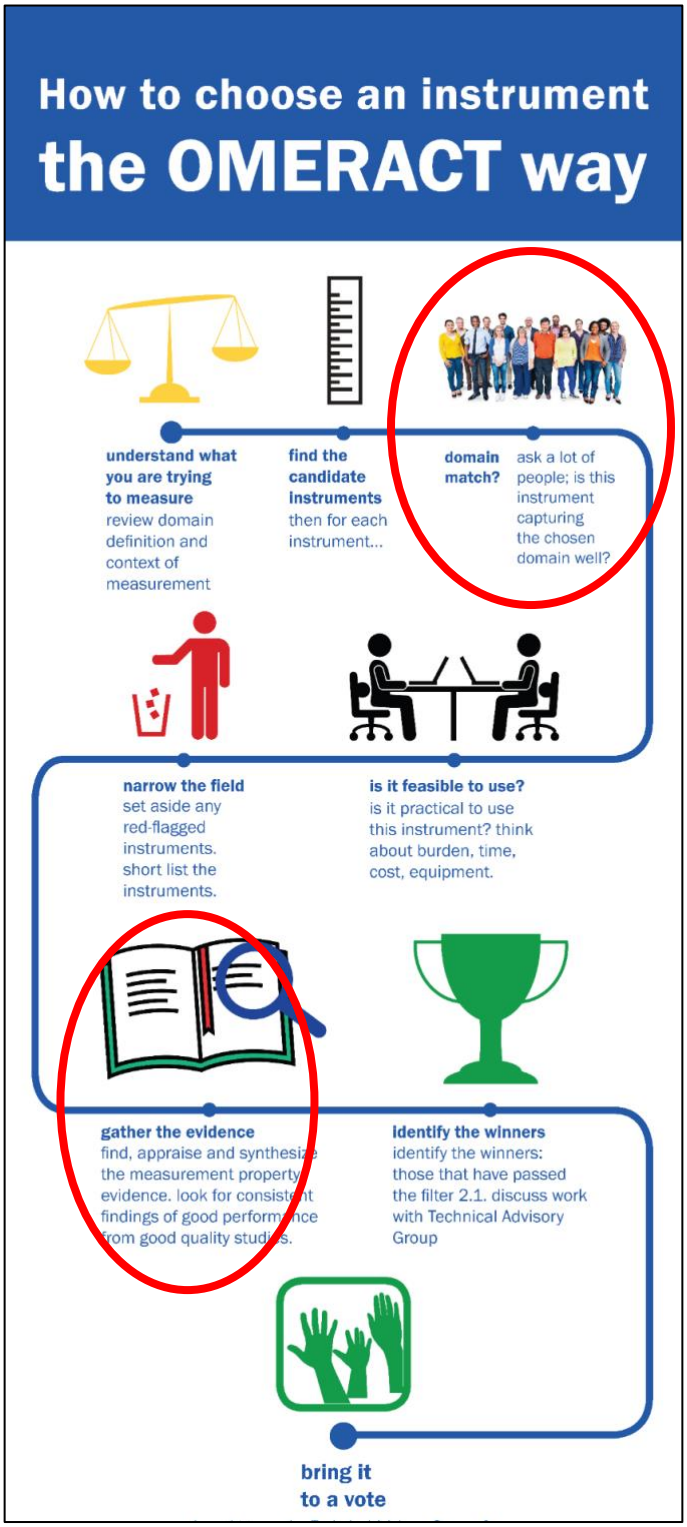
A.1. Sources of variability videos – Part A & Part B:

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

A.2. Lesson 2: Understand the sources of variability that impact your score

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

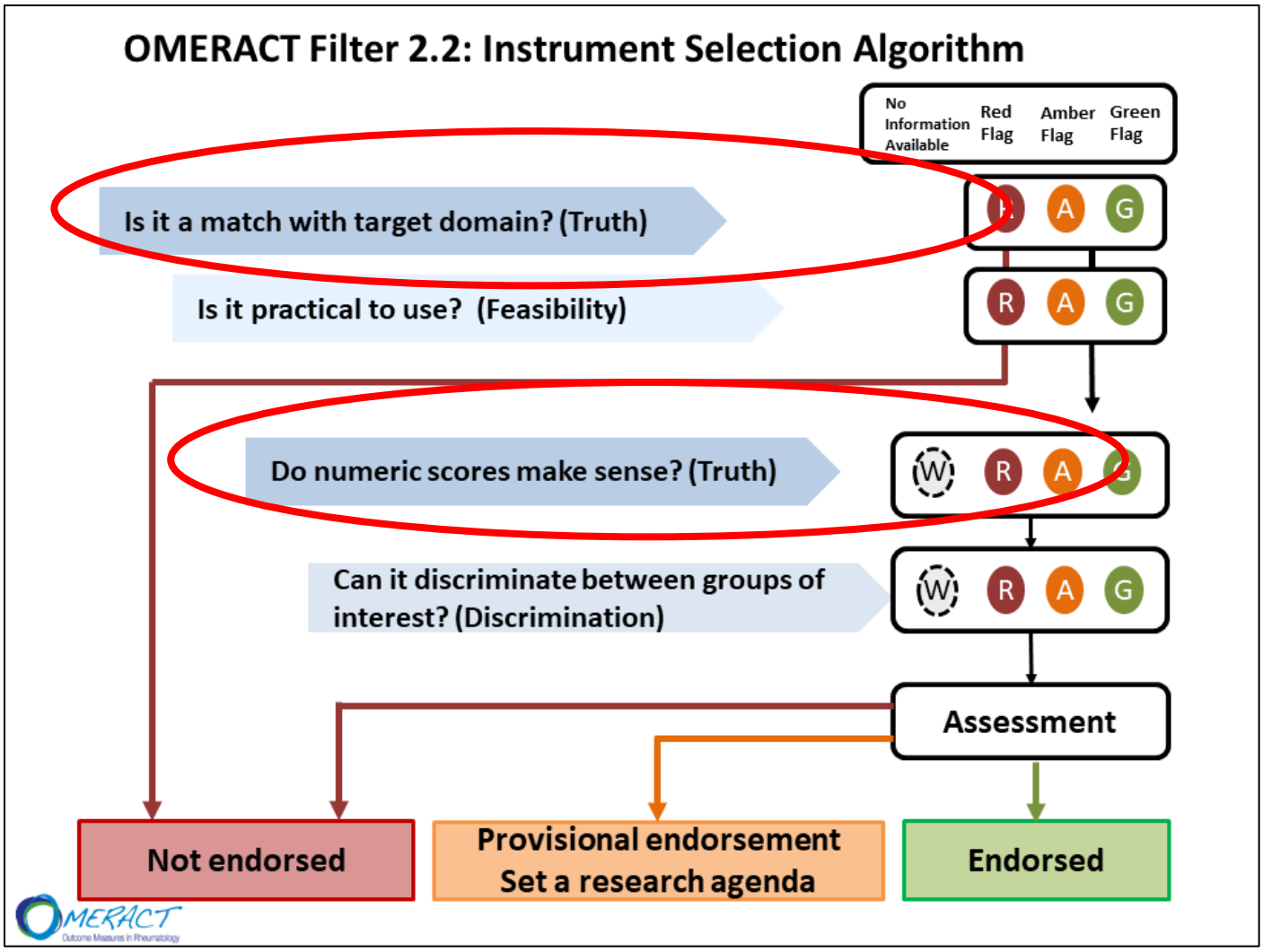
B. OMERACT Way



C. Master checklist for instrument selection: Steps 5 and 8-10

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

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



E. Where does testing sources of variability fit in the SOMP?

The sources of variability identified in the domain definition template can be assessed in two places:

1. At the domain match stage

2. When assessing the measurement property assessment of 'inter-method' reliability

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

F. Excerpt from OMERACT Handbook, Chapter 5, Instrument Selection (Page 14-17)

https://omeracthandbook.org/

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.

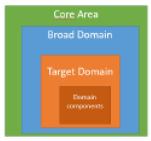
Essential to this assessment is the definition work done in the domain selection phase. Reviewing the domain definition template (see Figure 5.5) from the broad concept to the specific and focused target domain and its elemental components is important as an initial step to ensure that there is a match 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.

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

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".

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	○

G. Excerpt from Instrument selection workbook (pg. 7, 11-13, Appendix A)
https://omeracthandbook.org/workbooks

2.1 Target PICOC of Instrument

Define in detail the PICOC to which the instrument will apply [***you will take this from your Core Domain workbook***]

Domain Definition Report

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.

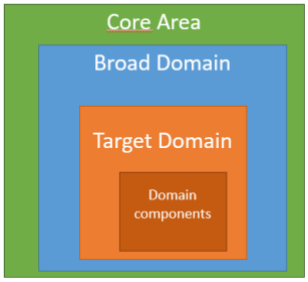
Working Group: _____

Target population _____

Intended use for this domain: _____ (e.g., RCT)

Intervention in trial: _____

Comparator in trial: _____



Core Area	*Life Impact*
Broad domain	*The general or 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 (https://doi.org/10.1136/annrheumdis-2020-217237), and work on lessons learned from*

	imaging outcomes (see video: Lesson 2: Understand the sources of variability that impact your score (https://omeract.org/instrument-selection/).

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)

To answer this question, there are four tasks to complete:

1. Survey of working group members about the content and domain match

2. Survey of patients and other key stakeholders about the content and domain match

3. Review of raw data for this instrument

4. Working group comes to a conclusion about match with target domain and content

5. 1 Survey of working group members about the content and domain match

Sample survey questions are provided in Appendix A. You can use any survey software to obtain this information.

Please provide a summary of your working group's input regarding the Domain Match of the selected instrument. Below are samples of the types of questions we need you to address; you can use these or similar questions but please provide a summary of your working group's responses at this level of detail. This example is geared towards PROs; we offer a table with suggestions of how to ask similar types of questions of imaging outcomes, other biomarkers and composite outcomes in Appendix A.

SUMMARY OF DOMAIN MATCH (IF NECESSARY, REPLACE & PROVIDE YOUR RESULTS HERE):

Instrument: Click here to enter text. **Date:**Click here to enter a date.

Question	Working Group's summary response	
Is this instrument (think about items, response, domain capture for PROs; for imaging, think about match with domain components) measuring what YOU want to measure? Are the items relevant to your concept, as experienced by your targeted patients' experiences and for the intended application? Consider sources of variability you identified, are any of those criteria that were considered in the definition of the domain? For example, is using assistive devices permitted in your concept of independence in ADL functioning? Or do you want to specify a particular time of day when you define your concept of pain intensity – night pain, or morning pain for example? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)

Have all important the elements of the target domain for this population, and intended application been included (consider breadth and depth needed)? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Is the instrument free of redundant, unnecessary, or potentially inappropriate or sensitive items? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Are the items phrased in a clear and understandable way? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Are the items written at a level that will be understood by the target population? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Are the instructions for completing items and selecting responses for the items clear? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Are the response options clear and appropriate for each item (consider match with the question, ordering of responses)? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)
Is the recall period in the instrument appropriate given the population, domain and intended application, i.e, over the past week, last 24 hours (if applicable)? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No ☐Not applicable	N (%) N (%) N (%)
Is the method of scoring appropriate (consider any weighted responses)? **Comments:** Click here to enter text.	☐Yes ☐Uncertain ☐No	N (%) N (%) N (%)

Final decision based on working group data(check one)**:**

☐**Good to go**

☐**some cautions but okay**

☐ **not right for this application**

.

Appendix A: Sample survey questions for surveying patients/other respondents on domain match and feasibility

The sample questions below are based on assessing domain match and feasibility of a PRO. They can be modified for use with other types of instruments using the considerations in this table:

PRO's	**Composites**	**Imaging Outcome Instrument (scoring system)**	**Other biomarkers (e.g. ESR, CRP)**
Match to domain definition	Is there a clear match with each domain to the target of the composite (i.e., disease activity?)	Does it capture the target domain?	Is this biomarker a good match to the domain definition?
Framing of the domain. Do the instructions for the instrument orient the respondent/ observer as to how to consider important sources of variability?	Does the questionnaire specify how to manage things like assistance, assistive devices, or coping strategies when answering?	Does the imaging technique description offer choices to avoid specific sources of variability? i.e., T2 weighting in MRI, or specific angle used for an Xray examination.	Does the instrument offer specific directions that help to avoid variability in scores? For example, blood pressure can vary by time of day, examiner, and environmental factors.
Do items cover the essential elements of the domain from detailed definition template?	Coverage of key elements of the target composite domain in the parts of the composite	Coverage of the elements of that domain	Is it capturing all the essential elements as described in the domain definition?
Response options	Scoring of each of the parts of the composite (remember inherent weighting given by the scaling of each domain)	Scoring of each element appropriate?	How is it quantified and is that standardized?
Weighting of items into score	How are elements weighted in the composite scoring system?	How the scoring system weights the elemental components in the score?	How is it scored compared to norms?

Feedback from <u>respondents</u> to instrument:

Is it a match with target domain? (Truth) & Is it practical to use? (Feasibility)

Instrument: Click here to enter text. **Date:**Click here to enter a date.

Match to Domain: Thinking about the content of the actual questions/items in the instrument, based on experience of this domain.	Respondents answer
Are the items in this instrument relevant to you and your experience? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Do you think there should be any additional items (i.e., were there things that were missed)? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Do you think that there should be any items taken out of the instrument? If yes, tell us why. Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Were there overlapping, sensitive, or embarrassing items? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Does the instrument overall reflect your experience of your [domain]? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Did you find that all the items were easy to read? If not, which items were not easy to read? Click here to enter text.	☐Yes ☐Uncertain ☐No
Did you feel that all the items were clear and understandable? Could you understand what all the questions were trying to ask? If not, which items did you feel were unclear? Click here to enter text.	☐Yes ☐Uncertain ☐No
Did you think that the response options were clear and understandable (i.e. did the possible answers match well with the questions)? If not, which items did you feel had a mismatched response scale?) Click here to enter text.	☐Yes ☐Uncertain ☐No
Were the instructions for answering the items clear? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Does the timing of the recall period seem reasonable to you (e.g. over the past week, last 24 hours) (if applicable)? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No ☐Not applicable
Feasibility: Questions about the practical considerations about this instrument.	Respondents answer
Was it easy enough to complete? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No

Did it take a reasonable amount of time to complete? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No
Did the format seem appropriate (how it looked on the page, font size, how items and responses were organized)?	☐Yes ☐Uncertain ☐No
Do you think there was too much equipment and training needed before you could be able to respond to this instrument? Comments: Click here to enter text.	☐Yes ☐Uncertain ☐No