
Zeeshan Butt, PhD  
David Cella, PhD  
Richard Gershon, PhD  
Elizabeth Rauch  
Kathleen Swantek, MLS

Bryce Reeve, PhD  
Jessica Lyons, MS
ISOQOL Scientific Advisory Task Force (SATF)

- Neil Aaronson, PhD  
  University of Amsterdam
- Sara Ahmed, PhD  
  McGill University
- Michael Brundage, MD  
  Queens University
- Peter Fayers, PhD  
  University of Aberdeen
- David Feeny, PhD  
  University of Alberta
- Joanne Greenhalgh, PhD  
  University of Leeds
- Ron Hays, PhD  
  University of California – Los Angeles
- Pamela Hinds, PhD  
  National Children’s Hospital, Wash. DC
- William Lenderking, PhD  
  United BioSource Corporation
- Lori McLeod, PhD  
  RTI – Health Solutions
- Carol Moinpour, PhD  
  Fred Hutchinson Cancer Research Center
- Dennis Revicki, PhD  
  United BioSource Corporation
- Carolyn Schwartz, ScD  
  DeltaQuest Found. / Tufts Univ. Med. Sch.
- Claire Snyder, PhD  
  Johns Hopkins University
- Caroline Terwee, PhD  
  VU University Medical Center
- Galina Velikova, MD, PhD  
  University of Leeds
- Albert Wu, MD  
  Johns Hopkins University
- Kathleen Wyrwich, PhD  
  United BioSource Corporation
• Comprehensive literature review

• Interact with ISOQOL SATF
• Field ISOQOL member survey.

Reports

<table>
<thead>
<tr>
<th>How we spent our winter vacation.</th>
</tr>
</thead>
</table>

Thou shall assess Reliability
Thou shall assess Validity
Literature Review

• Sources
  o ISOQOL membership
  o Pre-contract activity
  o Pubmed
  o PsycINFO
  o CINAHL

• Search Parameters
  o Adapted the methodological search filter developed by Terwee et al (2009) *Qual Life Res*
  o Focus was on consensus statements, guidelines, and evidence-based papers
  o Included population-, but not instrument-specific papers, if concepts were generalizable
  o Included unpublished and published sources

- 301 reviewed → 12 unique
- 24 reviewed → 12 unique
- 544 reviewed → 14 unique
- 172 reviewed → 22 unique
- 126 reviewed → 4 unique
ISOQOL Membership Survey

• Designed with ISOQOL SATF
• Sought consensus on draft recommendations presented today and explored more in depth on specific issues of debate.
• Study was approved by UNC IRB.
• Sent out on Feb. 20 to approximately 500 ISOQOL members with deadline of Feb. 29.
• ISOQOL SATF members reviewed responses to survey.
Challenges with writing recommendations for Minimum Standards.
Instructions on ISOQOL Survey

Please remember as you answer the questions in this survey that we are developing the minimum standards for the selection and design of a PRO measure for use in patient-centered outcomes research (PCOR).

That is, we are saying a PRO measure that does not meet the “minimum standard” should not be considered appropriate for the research study.
Guidance for Industry
Patient-Reported Outcome Measures:
Use in Medical Product Development
to Support Labeling Claims

The COSMIN checklist for a
set of studies on measurement
properties of measurement instruments:
a report of the COSMIN
Multicentre Study Group
Lidwine B. Mokkink · Caroline B. Terwee ·
Donald L. Patrick · Jordi Alonso · Paul W. S.
Dirk L. Knol · Lex M. Bouter · Henrica C. W.

Abstract

The field of health status and quality of life (QoL) has a
long theoretical framework, accepted methods, and defin- tion of 30 years. To identify health status and QoL as
precursors to creating an instrument library for clinical use I created an independently functioning Scientific
Advisory Committee on Measurement of Health Status
(SACMAHS) that has met every 2 years since 1997. We
defined a set of attributes and criteria to carry out our
work, and we have compiled a list of references that
are key to the field. This paper presents the results
of our work and provides a checklist for evaluating
QoL measures.

Key words: Health status, Item response theory, Validity

Introduction

The field of health assessment

The field of health status and quality of life (QoL)
measurement — as a formal discipline with a cer-

1 Neil Arronson, PhD, The Netherlands Cancer Institute, Amsterdam; Jordi Alonso, MD, Institut Municipal d’Investiga-
ció Media (IMIS-IMAS), Barcelona, Spain; Audrey Buran, PhD, The RAND Corporation, Santa Monica, CA; Kathleen
L. Lohr, PhD, RTI International, Research Triangle Park, NC, and Program on Health Outcomes, University of North Carolina at Chapel Hill; Donald L. Patrick, PhD, MSPH, Department of Health Services, University of Washington, Seattle; Edward Perrin, PhD, Department of Health Services, University of Washington, Seattle, Ruth E.K. See
MD Albert Einstein College of Medicine/Children’s Hospital Montefiore, Bronx, NY.
Attributes of a PRO Measure

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and Measurement Model</td>
</tr>
<tr>
<td>2</td>
<td>Reliability</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
</tr>
<tr>
<td>3a</td>
<td>- Content Validity</td>
</tr>
<tr>
<td>3b</td>
<td>- Construct Validity</td>
</tr>
<tr>
<td>3c</td>
<td>- Responsiveness</td>
</tr>
<tr>
<td>4</td>
<td>Interpretability of Scores</td>
</tr>
<tr>
<td>5</td>
<td>Translations</td>
</tr>
<tr>
<td>6</td>
<td>Patient and Administrator Burden</td>
</tr>
</tbody>
</table>
ISOQOL Sample Characteristics (n = 98)

- **Degrees**: 62% PhD; 17% MD; 45% Masters
- **Role**: 71% Academic; 19% Clinicians; 19% Industry Consultant; 8% Industry; 7% Government
- **Geographic**: 48% N. A. (85% US); 33% Europe; 9% Asia; 6% S.A.; 3% Australia; 1% Africa; 0% Antarctica
- **Psychometric training**: 81% mod – extensive; 16% little
- **Qualitative training**: 53% mod – extensive; 40% little
- **Competency**: 50% very competent; 39% competent; 8% somewhat competent
- **Average # of years HRQOL/PRO research**: 15 years (range 1 – 40 years)
Survey Question: Please provide your opinion on the following items regarding the minimum standard for a PRO measure:

RECOMMENDATION

• Required as a minimum standard
• Desirable but not required as a minimum standard
• Not required at all (not needed for a PRO measure)
• Not sure
• No opinion
Survey Question: Please provide your opinion on the following items regarding the **minimum standard** for a PRO measure’s conceptual and measurement model:

**RECOMMENDATION**

- Required as a minimum standard
- Desirable but not required as a minimum standard
- Not required at all (not needed for a PRO measure)
- Not sure
- No opinion

General rule:
> 50% accepted;
< 50% consider as “best practice”
1. Conceptual and Measurement Model

A PRO measure should have documentation defining and describing the concept(s) included and the intended population(s) for use. (91% required)

In addition, there should be documentation of how the concept(s) are organized into a measurement model, including evidence for the dimensionality of the measure, how items relate to each measured concept, and the relationship among concepts included in the PRO measure. (62% required)
2. Reliability

The reliability of a PRO measure should ideally be at or above 0.70 for group level comparisons.

- 55% agree;
- 35% → “no minimum level…it should be appropriately justified for the context of the proposed application”

Reliability for multi-item scales should include an assessment of internal consistency (81% required) and test-retest reliability; (44% required)

...and reliability for a single item measure should be assessed by test-retest reliability. (63% required)
3. Validity: How critical is each type of validity?

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Construct</th>
<th>Responsiveness</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cross-sectional studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Must have, or I wouldn’t use PROM</td>
<td>60%</td>
<td>49%</td>
<td>26%</td>
<td>8%</td>
</tr>
<tr>
<td>Would expect to have in most cases</td>
<td>35%</td>
<td>49%</td>
<td>52%</td>
<td>35%</td>
</tr>
<tr>
<td>Nice to have, but not critical</td>
<td>4%</td>
<td>1%</td>
<td>14%</td>
<td>44%</td>
</tr>
<tr>
<td>Not critical at all.</td>
<td></td>
<td></td>
<td>7%</td>
<td>11%</td>
</tr>
<tr>
<td><strong>Longitudinal studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Must have, or I wouldn’t use PROM</td>
<td>63%</td>
<td>45%</td>
<td>60%</td>
<td>11%</td>
</tr>
<tr>
<td>Would expect to have in most cases</td>
<td>33%</td>
<td>52%</td>
<td>38%</td>
<td>34%</td>
</tr>
<tr>
<td>Nice to have, but not critical</td>
<td>4%</td>
<td>3%</td>
<td>2%</td>
<td>41%</td>
</tr>
<tr>
<td>Not critical at all.</td>
<td></td>
<td></td>
<td></td>
<td>12%</td>
</tr>
</tbody>
</table>

Note: If a PRO Measure had cross-sectional data to support reliability and validity (content, construct); but no data on responsiveness: 64% would use PROM; 33% would require evidence of responsiveness before using it.
3a. Content Validity

A PRO measure should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application. This includes documentation of:

- qualitative and/or quantitative methods used to solicit and confirm attributes’ (i.e., concepts measured by the items) of the PRO relevant to the measurement application;

- the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, gender, socio-economic status, literacy level) with an emphasis on similarities or differences with respect to the target population;

- ways from which items were developed and/or sources from which items were derived, modified, and prioritized during the PRO measure development process; and

- justification for the recall period for the measurement application.
3a. Content Validity (part i)

A PRO measure should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PRO measure relevant and comprehensive for the concept, population, and aim of the measurement application. (78% required)
3a. Content Validity (part ii)

This includes documentation of:

1. qualitative and/or quantitative methods used to solicit and confirm attributes (i.e., concepts measured by the items) of the PRO relevant to the measurement application; (53% required)

2. the characteristics of participants included in the evaluation (e.g., race/ethnicity, culture, age, gender, socio-economic status, literacy level) with an emphasis on similarities or differences with respect to the target population; (53% required)

3. ways from which items were developed and/or sources from which items were derived, modified, and prioritized during the PRO measure development process; (47% required)

4. justification for the recall period for the measurement application. (42% required)
3b. Construct Validity

A PRO measure should have evidence supporting its construct validity, including documentation of

- empirical findings that support predefined hypotheses on the expected associations among measures similar or dissimilar to the measured PRO, (55% required)

- or expected differences in scores on that PRO measure between “known” groups. (41% required)
3c. Responsiveness

A PRO measure for use in longitudinal research study should have evidence of responsiveness, including empirical evidence of changes in scores consistent with predefined hypotheses regarding changes in the target population for the research application. (57% required)

**Note:** However, there may be circumstances in which a PRO measure with cross-sectional data to support reliability and validity (content, construct), but no data on responsiveness, could be used (64% supporting such use).
4. Interpretability of Scores

A PRO measure should have documentation to support interpretation of scores, including:

• what low and high scores represent for the measured concept; (65% required)

• representative mean(s) and standard deviation(s) in the reference population; (40% required)

• guidance on the minimally important difference in scores between groups and/or over time that can be considered meaningful from the patient and/or clinical perspective. (23% required)
5. Translations

A PRO measure translated to one or more languages should have evidence of the equivalence of measurement properties for translated versions, allowing comparison or combination of data across language forms. *(48% required)*

<table>
<thead>
<tr>
<th></th>
<th>Must have</th>
<th>Expect in most cases</th>
<th>Nice to have, but not critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative evidence</td>
<td>42%</td>
<td>39%</td>
<td>18%</td>
</tr>
<tr>
<td>Quantitative evidence</td>
<td>23%</td>
<td>43%</td>
<td>30%</td>
</tr>
</tbody>
</table>

This includes documentation of:

1) background and experience of the persons involved in the translation; *(42% required)*

2) methods used to translate and evaluate the PRO measure in each language; *(81% required)*

3) extent of harmonization across different language versions. *(38% required)*
5. Translations

A PRO measure translated to one or more languages should have documentation of the methods used to translate and evaluate the PRO measure in each language. Studies should at least include evidence from qualitative methods (e.g., cognitive testing) to evaluate the translations.
6. Patient and Administrator Burden

A PRO measure must not be overly burdensome for patients or administrators. The length of the PRO measure should be considered in the context of other PRO measures included in the assessment, the frequency of PRO data collection, and the characteristics of the study population.

The literacy demand of the items in the PRO measure should usually be at a 6th grade education level or lower.

- 4th grade: 6%
- 6th grade: 23%
- 8th grade: 6%
- 44% endorsed “There should be no minimum requirement for the literacy level of the PRO measure; however, it should be appropriately justified for the context of its proposed application.”
6. Patient and Administrator Burden

A PRO measure must not be overly burdensome for patients or administrators. The length of the PRO measure should be considered in the context of other PRO measures included in the assessment, the frequency of PRO data collection, and the characteristics of the study population.

The literacy demand of the items in the PRO measure should usually be at a 6th grade education level or lower; however, it should be appropriately justified for the context of the proposed application.
How can these standards be used by a PRO Measure Developer or Investigator?

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and Measurement Model</td>
</tr>
<tr>
<td>2</td>
<td>Reliability</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
</tr>
<tr>
<td>3a</td>
<td>- Content Validity</td>
</tr>
<tr>
<td>3b</td>
<td>- Construct Validity</td>
</tr>
<tr>
<td>3c</td>
<td>- Responsiveness</td>
</tr>
<tr>
<td>4</td>
<td>Interpretability of Scores</td>
</tr>
<tr>
<td>5</td>
<td>Translations</td>
</tr>
<tr>
<td>6</td>
<td>Patient and Administrator Burden</td>
</tr>
</tbody>
</table>
How can these standards be used by a PRO Measure Developer?

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Method</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and Measurement Model</td>
<td>PRO defined and items generated through expert panel.</td>
<td>This domain includes 4 attributes measured by 20 items.</td>
</tr>
<tr>
<td>2</td>
<td>Reliability</td>
<td>Internal Consistency (Cronbach’s alpha) at baseline; Test-retest between Baseline &amp; 1 day later.</td>
<td>Internal consistency (n = 258) $\alpha = .84$; Test-retest (n = 50) $r = .72$</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>- Content Validity</td>
<td>Literature review and focus groups to identify sub-domains of concept. Cognitive testing to evaluate relevance of items and modify problematic items.</td>
<td>2 focus groups (n = 17) identified 4 attributes. Consistent with theory of Awesomeness. Cog. Testing (n = 8, round 1; n=6, round 2) 4 items modified.</td>
</tr>
<tr>
<td>3b</td>
<td>- Construct Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td>- Responsiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Interpretability of Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Translations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Patient and Administrator Burden</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Question:

How can these standards be used by an Investigator? (1)

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Evidence?</th>
<th>Relevance &amp; needs for current research application</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and Measurement Model</td>
<td>Definition…. ; 4 Attributes with 20 items. In Smith et al. 2011.</td>
<td>PRO is highly prevalent and bothersome in our population.</td>
</tr>
<tr>
<td>2</td>
<td>Reliability</td>
<td>Internal consistency (n = 258) $\alpha = .84$; Test-retest (n = 50) $r = .72$. In Smith et al. 2011.</td>
<td>Acceptable for prospective study among 2 arms.</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
<td>Validation study in Thomas et al was older than our target pop.</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>Content Validity</td>
<td>2 focus groups in older men with prostate cancer (Thomas et al. 2012)</td>
<td>Need to conduct 2 focus groups with young prostate cancer patients to confirm relevant attributes.</td>
</tr>
<tr>
<td>3b</td>
<td>Construct Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td>Responsiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Interpretability of Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Translations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Patient and Administrator Burden</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## How can these standards be used by an Investigator? (2)

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Evidence for Instrument A</th>
<th>Evidence for Instrument B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Conceptual and Measurement Model</td>
<td>Definition…. ; 4 Attributes with 20 items.</td>
<td>Definition….; 2 Attributes with 16 items.</td>
</tr>
<tr>
<td>2</td>
<td>Reliability</td>
<td>Internal consistency (n = 258) $\alpha = .84$; Test-retest (n = 50) $r = .72$.</td>
<td>Internal consistency (n = 312) $\alpha = .91$; Test-retest (n = 78) $r = .74$.</td>
</tr>
<tr>
<td>3</td>
<td>Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>- Content Validity</td>
<td>2 focus groups in older men with prostate cancer.</td>
<td>3 focus groups in younger and older prostate cancer population.</td>
</tr>
<tr>
<td>3b</td>
<td>- Construct Validity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td>- Responsiveness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Interpretability of Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Translations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Patient and Administrator Burden</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Beyond the PCORI Contract Period

• Develop “Ideal” or “Best Practices” standards...

• PRO Measure Maturity (or Stepped) Model
  – From PROMIS, “… describes the stages of instrument scientific development from conceptualization through evidence of psychometric properties in multiple diverse populations.”

• Standards for PRO Measures to be used in the healthcare setting.