What are the most effective ways to measure patient health outcomes of primary health care integration through PROM (Patient Reported Outcome Measurement) instruments?

Stirling Bryan\textsuperscript{1,2}, James Broesch\textsuperscript{1}, Kacey Dalzell\textsuperscript{1}, Jennifer Davis\textsuperscript{1,2}, Martin Dawes\textsuperscript{3}, Mary M. Doyle-Waters\textsuperscript{1}, Steven Lewis\textsuperscript{4}, Kim McGrai\textsuperscript{2,5}, Margaret J. McGregor\textsuperscript{1,2}, Janice M. Murphy\textsuperscript{6}, Rick Sawatzky\textsuperscript{7,8}

1. Centre for Clinical Epidemiology & Evaluation (C2E2), Vancouver Coastal Health Research Institute (VCHRI)
2. The University of British Columbia, School of Population & Public Health
3. The University of British Columbia, Department of Family Practice, Faculty of Medicine
4. Access Consulting Ltd.
5. The University of British Columbia, Centre for Health Service & Policy Research (CHSPR)
6. Research Consultant, Balfour, BC
7. Trinity Western University, School of Nursing
8. Centre for Health Evaluation and Outcome Sciences, Providence Healthcare

April 2013

This study was supported by the Canadian Institutes of Health Research.
# Table of Contents

Table of Contents .......................................................................................................................... 2
List of Tables .................................................................................................................................. 4
List of Figures ................................................................................................................................. 5
Abbreviations ................................................................................................................................. 5
Executive Summary ......................................................................................................................... 6
  Project components ....................................................................................................................... 6
  Descriptive overview .................................................................................................................... 7
  Psychometric and decision making review ................................................................................... 7
  Workshop and recommendations ................................................................................................. 8
Introduction and Background ......................................................................................................... 9
Research Question and Study Components .................................................................................. 11
Generating the Long and Short Lists of Generic PROMs (Project components 1 and 2) .......... 14
Descriptive Overview of Short-listed PROMs (Project component 3) ........................................ 17
Methods ......................................................................................................................................... 17
  Domains ...................................................................................................................................... 17
  Response Options ......................................................................................................................... 17
  Burden on Respondent .................................................................................................................. 17
  Readability ................................................................................................................................. 18
  Recall Period .............................................................................................................................. 18
  Reported Modes and Methods of Administration ....................................................................... 18
  Translations ............................................................................................................................... 18
  Instrument Cost ........................................................................................................................... 18
Results ............................................................................................................................................ 18
  Domains ...................................................................................................................................... 18
  Response Options ......................................................................................................................... 19
  Burden on Respondent .................................................................................................................. 20
  Readability ................................................................................................................................. 20
  Recall Period .............................................................................................................................. 21
  Reported Modes and Methods of Administration ....................................................................... 21
  Translations ............................................................................................................................... 22
List of Tables

Table 1: Overview of results from psychometric review .......................................................... 8
Table 2. Response options ........................................................................................................ 19
Table 3. Respondent burden ..................................................................................................... 20
Table 4. Readability .................................................................................................................. 20
Table 5. Recall period ............................................................................................................... 21
Table 6. Administration methods and modes ............................................................................ 22
Table 7. Translations ................................................................................................................. 22
Table 8. Instrument cost ............................................................................................................ 23
Table 9. Overview of results from psychometric review ......................................................... 28
Table 10. Utility scores, population norms, minimally important differences ......................... 32
Table 11. Use of instruments in primary care settings ............................................................... 34
Table 12. Examples of use of instruments in other jurisdictions ............................................... 35
Table 13. Cited references in the Web of Science ..................................................................... 38
Table 14. Patient reported outcomes measurement group bibliography search ..................... 40
Table 15. Total citations for the years 2007 to 2012 ............................................................... 42
Table 16. Instrument details ...................................................................................................... 53
Table 17. Age groups instruments have been adapted for ....................................................... 56
Table 18. Legend for the summary of the psychometric evidence ............................................. 57
Table 19. Summary of the psychometric evidence of the “A-List” (Top three instruments selected by January 14, 2013 workshop participants) ............................................................. 57
Table 20. Summary of the psychometric evidence of the “B-List” (Three instruments eliminated from the final list in the second cut by January 14, 2013 workshop participants) ............... 59
Table 21. Summary of the psychometric evidence of the “C-List” (First two instruments to be eliminated from the final list in the first cut by January 14, 2013 workshop participants) ........ 61
Table 22. Quality assessment of the articles reviewed .............................................................. 63
List of Figures

Figure 1. Generating the short list of PROM instruments ..........................................................15
Figure 2. Domain coverage* ..................................................................................................19
Figure 3. Results of search for reviews of psychometric properties ........................................52

Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>Assessment of Quality of Life</td>
</tr>
<tr>
<td>CIHR</td>
<td>Canadian Institutes of Health Research</td>
</tr>
<tr>
<td>COSMIN</td>
<td>Consensus-based Standards for the Selection of Health Measurement Instruments</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Euroqol EQ-5D</td>
</tr>
<tr>
<td>HUI®: HUI3</td>
<td>Health Utilities Index</td>
</tr>
<tr>
<td>NHP</td>
<td>Nottingham Health Profile</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>PROMIS: GHS</td>
<td>Patient Reported Outcomes Measurement Information System: Global Health Scale</td>
</tr>
<tr>
<td>PROQOLID</td>
<td>Patient-Reported Outcome and Quality of Life Instruments Database</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>Quality of Well-Being scale - Self-Administered</td>
</tr>
<tr>
<td>SF-12® / SF-12v2™</td>
<td>SF-12® Health Survey and SF-12v2™ Health Survey</td>
</tr>
<tr>
<td>SF-36® / SF-36v2™</td>
<td>SF-36® Health Survey and SF-36v2™ Health Survey</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>World Health Organization Quality of Life assessment instrument - BREF</td>
</tr>
</tbody>
</table>
Executive Summary

This project was funded by the Canadian Institutes of Health Research (CIHR) as an Expedited Knowledge Synthesis (EKS). The policy topic for the work was identified by the British Columbia (BC) Ministry of Health, and so the requirement of the grant was for the research team to work in partnership with policy colleagues at the Ministry and the Michael Smith Foundation for Health Research (MSFHR).

The research question stated in the CIHR funding announcement was: ‘What are the most effective ways to measure patient health outcomes of primary health care integration through Patient Reported Outcome Measurement Survey instruments?’

Discussion with Ministry colleagues clarified the research focus, indicating that the interest was in generic and not condition-specific instruments and in the potential utility of PROMs in the context of health care system change, not just ‘integration’. It was agreed that the project’s scope should exclude residential or institutional care. It was also agreed that instruments solely measuring patients’ experience of care (PREMS) would be excluded.

Given the complexity and multi-faceted nature of the project, with tight timelines for delivery, we used a ‘rapid review’ approach that specifically integrated knowledge translation approaches to ensure that the outcomes of the review were relevant to the target audience.

Project components

The project had 6 main components.

(1) Long-list of generic PROMs: a comprehensive long-list of all generic PROM instruments

(2) Short-list of generic PROMs: To include truly generic, quantitative measures designed for adult populations and with high recent citation counts

(3) Descriptive overview of short-listed PROMs: To include official translations, respondent burden including required literacy/reading level, cost for using the instrument, dimension coverage

(4) Review of PROM instrument ‘performance’: Psychometric (e.g., reliability, validity, responsiveness and interpretability) and decision-making (e.g., norm reference sets, utility/preference scoring algorithm, evidence of clinically relevant thresholds)

(5) Additional information: examples of use in a primary and community care context and PROM-related activity in other jurisdictions

(6) Workshop and recommendations: To review the evidence and identify the ‘preferred’ instrument (or instruments) for use in BC primary and community care reform
At the end of component 2, the short-list comprised:

- Assessment of Quality of Life (AQoL-8D)
- EuroQol EQ-5D
- HowsYourHealth
- Health Utilities Index (HUI)
- Nottingham Health Profile (NHP)
- PROMIS/GHS
- Short-Form 36 (SF-36) and SF-12
- Quality of Wellbeing Scale (QWB-SA)
- World Health Organization Quality of Life Instrument (WHOQoL-BREF)

**Descriptive overview**

In terms of domain coverage, mental and physical health domains were represented in all short-listed instruments but not all instruments included questions about social health and general health.

In terms of respondent burden, the EQ-5D, SF-12 and PROMIS/GHS have the fewest questions and take the least time to complete. However, the readability scores were most favourable for AQoL-8D, SF-36, and QWB-SA. Translation work has been most extensive for the EQ-5D, SF-12 and SF-36 – these instruments are available in most of the languages used frequently in BC.

Many of the instruments are available for use for free: AQoL, NHP, WHOQoL-BREF and the PROMIS/GHS. Although the SF-36 is not free, both the RAND and Veterans Administration versions of the SF-36 are free for use.

**Psychometric and decision making review**

The literature search was for published and grey review articles that discussed the psychometric properties of at least one of the short-listed instruments, except for HowsYourHealth (there were no articles that addressed the psychometric properties of this instrument). Searches were undertaken using bibliographic databases (MEDLINE and Embase) supplemented with general web searches. A formal review approach was adopted, including study selection, data extraction and quality assessment. Data synthesis involved a scoring approach, proposed by the COSMIN-initiative, for systematic reviews of the measurement properties of instruments. We considered aspects of each instrument’s reliability, validity and responsiveness, following strict guidelines on scoring in each of these categories. We included two additional categories: generalizability and comparison with other PROMs instruments.

Twenty-one of the 22 articles provided information about the psychometric properties of the candidate PROM instruments. An overview of the results from the psychometric review is provided in Table 1 below. A general conclusion is that the SF-36 performed particularly well across most psychometric dimensions and PROMIS was also a strong instrument although the evidence base was smaller.
### Table 1: Overview of results from psychometric review

<table>
<thead>
<tr>
<th></th>
<th>AQoL</th>
<th>EQ-5D</th>
<th>SF-36</th>
<th>HUI</th>
<th>NHP</th>
<th>QWB-SA</th>
<th>WHOQoL-BREF</th>
<th>PROMIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal consistency</strong></td>
<td>+</td>
<td>n/a</td>
<td>+++</td>
<td>?</td>
<td>+/-</td>
<td>n/a</td>
<td>+/-</td>
<td>+++</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>?</td>
<td>+/-</td>
<td>+++</td>
<td>+/-</td>
<td>+/-</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td><strong>Content validity</strong></td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
<td>?</td>
<td>+/-</td>
</tr>
<tr>
<td><strong>Construct validity</strong></td>
<td>?</td>
<td>?</td>
<td>+++</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td><strong>Cross-cultural validity</strong></td>
<td>?</td>
<td>+</td>
<td>+/-</td>
<td>+</td>
<td>+/-</td>
<td>?</td>
<td>+++</td>
<td>?</td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>+</td>
<td>--</td>
<td>+++</td>
<td>+/-</td>
<td>+/-</td>
<td>-</td>
<td>+/-</td>
<td>++</td>
</tr>
</tbody>
</table>

Note: + = Positive Evidence; - = Negative Evidence; +/- = Conflicting Evidence; ? = Unknown/Not Reported

In terms of decision making criteria, although most of the instruments can provide utility scores, only EQ-5D and HUI3 have utility scores derived from a Canadian sample. Likewise, all the instruments provide population norms, but only the EQ-5D, SF-12/SF-36 and HUI3 supply Canadian population norms.

**Workshop and recommendations**

The project culminated with a workshop, held at the BC Ministry of Health and involving a wide range of stakeholders. The primary objectives for the whole workshop were:

- To share the details of the review and evidence synthesis work undertaken by the research team
- To have reflection and deliberation on the PROM instruments and their potential use in IPCC projects
- To come to a consensus on which PROM instrument(s) should be recommended for use in BC

There were two rounds of voting and discussion at the Workshop which occurred after information was presented on the review and participants had an opportunity to visually review the items within each measure. The final votes indicated a strong preference for two instruments, PROMIS (41 votes) and SF (36 votes), and a desire to keep EQ-5D (18 votes) as a back-up.
Introduction and Background

The project described in this report was funded by the Canadian Institutes of Health Research (CIHR) as an Expedited Knowledge Synthesis (EKS) project [1]. The policy topic for the work was identified by the British Columbia (BC) Ministry of Health, and so the requirement of the grant was for the research team to work in partnership with policy colleagues at the Ministry and the Michael Smith Foundation for Health Research (MSFHR). The culmination of the work was the end-of-project workshop, held at the Ministry of Health but involving a wide range of stakeholder partners. This report provides a description of all components of the project.

Why outcome measurement? The Institute for Healthcare Improvement (IHI) promotes health care system redesign that simultaneously accomplishes three key objectives (the “Triple Aim” [2]): improving population health, enhancing patient care experience (including quality, appropriateness and satisfaction), and controlling the rate of cost increases. This framework has been widely adopted as a guide for judging the success of system reform in Canada and is at the core of our interest in outcome measurement [3,4].

Making the Triple Aim operational in the health care system requires routine gathering of information on all three components. This is an agenda of “measurement to support improvement” in a dynamic health care system. Our focus in this project was on the health outcomes piece of the Triple Aim approach [2]. Recognition of the importance of outcomes is not new: ‘Outcomes, by and large, remain the ultimate validators of the effectiveness and quality of medical care’ [5, p169]. Our roles as health service analysts and stewards of the health care system demand that we have information on outcomes and investigate what systems or structures produce them.

Why patient-reported measures? Members of our research team have a long-standing commitment to giving voice to patients and the public in health care reform and policy [6,7]. This commitment is driven by recognition that, in the vast majority of circumstances, individuals are the best judges of their own health and wellbeing. There is widespread international interest in the use of patient reported outcome measures (PROMs) to monitor the effectiveness of health care services and interventions. Regulatory agencies, including the U.S. Food and Drug Administration and the United Kingdom’s (UK) National Health Service, are increasingly requiring the use of PROMs [8]. Finally, while data on death and hospitalization are commonly used, and widely available outcomes for evaluating health care, they fail to capture many other important aspects of the lives of people. The use of PROMs, in combination with these standard metrics, provides a fuller picture of the effects of the health care system on patients’ lives.

What outcome measures are available? In Canada, we do have some routine information on outcomes, such as mortality statistics and the suite of health indicators developed by the Canadian Institute for Health Information (CIHI) and Statistics Canada. This is good but limited information, often reflecting small samples and a focus on ‘failures’ such as hospital readmissions. Measurement of ‘success’ in terms of improvements in patients’ self-reported
health status or health-related quality of life (i.e., PROMs) is virtually non-existent in Canadian health care.

Though absent in routine practice in Canada, PROMs have been used widely in clinical trials and other research settings [9,10]. There now exist numerous patient reported outcome instruments, many of which have been well validated, some using Canadian populations [8]. PROMs are designed to measure either ‘general’ health status (i.e., generic PROMs [11,12,13]) or health status relating to a specific condition (i.e., condition-specific PROMs [14,15]). The focus here, as requested by the BC Ministry of Health, was solely on generic PROMs.

The main advantage of generic instruments is that they are suitable for use across a broad range of health problems. They can be used for comparisons of treatments for different patient groups and are suitable for use with healthy populations to generate normative data. Their broad scope means that they have potential to capture the influence of co-morbidity on health, as well as unexpected positive or negative effects of an intervention. The flip side is that their broad applicability means that some level of detail or sensitivity is sacrificed, which may limit the relevance of generic instruments when applied to a specific patient population. By definition, generic instruments are potentially less responsive to clinically important changes in health when compared to condition-specific measures [16]. In terms of the system it may be best to capture both generic information to get a picture of the whole, as well as some condition-specific information to review care of people with single diseases. The latter is further complicated by the increasing prevalence of multi-morbidity and requiring a broader generic capture of health status in someone with three or four individual conditions.
Research Question and Study Components

The research question stated in the CIHR funding announcement was:

‘What are the most effective ways to measure patient health outcomes of primary health care integration through Patient Reported Outcome Measurement Survey instruments?’

Discussion with Ministry colleagues clarified the research focus, indicating that the interest was in generic and not condition-specific instruments and in the potential utility of PROMs in the context of health care system change, not just ‘integration’.

The primary health care focus was also clarified and the CIHR definition of community-based primary health care (CBPHC) was adopted for the project. In broad terms, this covers primary prevention and primary care services within the community, including the management of chronic and episodic illness. It also involves the coordination and provision of integrated care provided by a range of health providers [17].

In our discussions with the Ministry, it was agreed that the project’s scope should include situations where the patient makes a direct contact with a provider in the community but should exclude residential or institutional care. It was also agreed that instruments solely measuring patients’ experience of care (PREMS) would be excluded.

The research question for this review required information from methodology sources, as well as evidence from clinical research and policy sources. Given the complexity and multi-faceted nature of the project, with tight timelines for delivery, we used a ‘rapid review’ approach that specifically integrated knowledge translation approaches to ensure that the outcomes of the review were relevant to the target audiences [18,19]. Regular working meetings with policy colleagues and advisors were held to provide opportunities for course correction, and ensure direct policy-engagement in the research process.

The project had 6 main components.

(1) Long-list of generic PROMs

The first task was to generate a comprehensive long-list of all generic PROM instruments, as a starting point for instrument selection.

(2) Short-list of generic PROMs

A two-stage short-listing process was adopted. The first stage involved filtering the long-list to ensure we only included instruments that are truly generic and self-report, assess health-related quality of life (i.e., not a single domain such as physical functioning, and not satisfaction or experience), were developed for an adult population, and are quantitative in nature. The second stage was to filter further on the basis of perceived utilization, which
was assessed via a citation reference search to reduce the list to a manageable number, given the project constraints. A decision was made to limit to a maximum of 10 instruments.

(3) Descriptive overview of short-listed PROMs

For all short-listed PROMs, some basic descriptive information relating to the practical use of the instruments was seen as necessary to inform selection. This information related to availability of official translations, respondent burden including required literacy/reading level, cost for using the instrument, dimension coverage, etc.

(4) Review of PROM instrument ‘performance’

For the short-listed instruments, to establish instrument ‘performance’, a formal review was undertaken to identify evidence on psychometric properties. This was supplemented with information relating to decision-making aspects of the short listed instruments.

(a) The psychometric criteria adopted in this work were based on the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) [20]. This includes reliability, validity, responsiveness and interpretability.

(b) Decision-making criteria were developed to consider instrument appropriateness for policy, management and clinical practice. The criteria included: norm reference sets, utility/preference scoring algorithm, evidence of clinically relevant thresholds, etc.

For the short-listed instruments, to establish instrument ‘performance’, a formal review was undertaken to identify evidence on psychometric properties. This was supplemented with information relating to decision-making aspects of the short listed instruments.

(5) Additional information

For the short-listed PROM instruments, additional information was identified, including examples of use in a primary and community care context (with a focus on primary care reform or practice change) and PROM-related activity in other jurisdictions (with a particular focus on other Canadian jurisdictions but also the UK and United States (US).

(6) Workshop and recommendations

The project culminated with a workshop, held at the BC Ministry of Health and involving a wide range of stakeholders. The primary objective of the workshop was two-fold: to review the evidence and information gathered as part of the project, and to identify the ‘preferred’ instrument (or instruments) for use in the context of primary and community care reform in BC. The outputs of the workshop represent the final recommendation on instrument selection.

The report is structured to provide an overview of methods and results for each component.
A project reference group was created, mainly comprising local stakeholders to supplement input from the BC Ministry of Health representatives who oversaw the project. The reference group was engaged to ensure that the refined research questions at the outset were aligned with local needs, to support on any specific topics where their input was required, and at the end of the synthesis process, to ensure that the recommendations made sense in the local contexts. In addition to Ministry colleagues, the following stakeholders acted in this role: Janet Joy (Vancouver Coastal HA), Connie Davis (Patients as Partners), Yoel Robens-Paradise (Lower Mainland Information Management), John Wasson (Dartmouth, US) and Garey Mazowita (Providence Health Care).

(See Appendix 9 for list of local stakeholders who participated in the PROMS Workshop on January 14, 2013).
Generating the Long and Short Lists of Generic PROMs (Project components 1 and 2)

Three approaches were used to generate a comprehensive long list of generic PROM instruments: a review of previously identified reviews, structured new searches, and consultation with PROM experts. In the protocol three reviews considered relevant to this project were referenced [8,21,22].

The structured searches focused on databases that indexed patient reported outcomes. An initial search was undertaken of the Health and Psychosocial Instruments (HaPI) database and, while it provided information on a number of relevant instruments, it did not easily group patient reported outcomes together. An Internet search for relevant databases uncovered two valuable sources:

- The Patient Reported Outcomes Measurement Group (phi.uhce.ox.ac.uk)
  This site, current to 2005, enables a search for generic PRO measures.

- The Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID)
  (www.proqolid.org)
  This site, developed by the MAPI Research Trust, includes measures that have been captured through “information taken from scientific literature and/or validated by the authors of the instruments” [23]. The database includes 115 generic measures.

Data were extracted from both sources but our primary focus was the more comprehensive and current PROQOLID database.

Figure 1 provides an overview of the short-listing process. KD developed a spreadsheet and collected information relating to all 115 generic measures listed on PROQOLID. Of those, 43 measures were selected on the basis of being truly generic and designed for an adult population, allowing patient self-administration, and assessing health-related quality of life or health status.

Access to the full PROQOLID database is only available through subscription. The fee was paid (a cost met from the grant) in order to gather further information for filtering. The research team reviewed the information on the 43 measures and also considered language availability, number of items, recall period and availability of other versions. Each measure was reviewed for appropriateness within the review based on the “objective” reported for the measure. If the focus was too narrow, such as on a single surgical intervention or for use in chronic populations the instrument was removed from consideration. Only broad population-based, generic outcome measures were selected for inclusion. After consideration of these factors the list was reduced to 25.

As part of our engagement with project advisors and policy colleagues, three additional instruments, not included in PROQOLID, were added to our list: HowsYourHealth, the Patient
Reported Outcomes Measurement System (PROMIS) and RAND-36. The total number of eligible instruments was 28.

**Figure 1. Generating the short list of PROM instruments**

The next filtering stage involved reviewing the perceived utilization of the 28 measures through a citation reference search to reduce the list to a manageable number, given the project’s constraints. A decision was made to limit to a maximum of 10 instruments.

This was achieved by examining a sample of the published peer-review literature to ascertain the ‘impact’ of each instrument since its development date. Impact was judged crudely in terms of citations. The Web of Science covers relevant disciplines and enables tracking of the number of times an article has been cited by other researchers. Creating a table of the number of cited references provided the team with some knowledge of an instrument’s use. The 28 generic instruments were included in the cited reference search.

PROQOLID gives details of original references for each instrument; these references were used in the cited reference search. The instrument’s title and acronym were also searched. All references that mentioned one of the instruments in the title, or were an original reference and had been cited, were eligible for inclusion in the total number of cites (see Appendix 1). Each instrument was also searched in the list of measures and as a keyword in the bibliography for...
The Patient-Reported Outcomes Measurement Group (phi.uhce.ox.ac.uk). The number of records that mentioned each instrument was recorded (see Appendix 2).

In order to have a feel for current practice in PROMs research, a decision was made to count cites in the last 6 years, in addition to total citation counts. Some instruments that were in high use are now in less demand (See Appendix 3).

The team reviewed the data gathered on all 28 instruments, including the overall number of citations and the change in citation rate in recent years. This resulted in 19 measures being removed from the list. The short-list comprised:

- Assessment of Quality of Life (AQoL)
- EuroQol EQ-5D
- HowsYourHealth
- Health Utilities Index (HUI)
- Nottingham Health Profile (NHP)
- PROMIS
- Short-Form 36 (SF-36) and SF-12
- Quality of Wellbeing Scale (QWB)
- World Health Organization Quality of Life Instrument (WHOQoL)

Most of the instruments listed above have a number of different versions. In order for the review work to be focused, the general rule was for only one instrument version to be considered. The selection of instrument version was guided by the following criteria.

- The instrument version must be patient self-report. For example, the version of the self-administration version of the QWB was used (QWB-SA).
- The instrument must be a discrete set of questions and not a full measurement system. This had implications for PROMIS, given its computer-adaptive nature drawing on a very large item-bank, such that each administration can involve different items. Therefore, our focus for this work was on the general health scale for PROMIS, known as PROMIS-GHS.
- Where instruments have versions that differ in terms of length then our focus was on the most feasible and hence shorter version. For example, the WHOQoL measure has both long (100 items) and short (26 items) versions; the shorter version is known as WHOQoL-BREF.

In the case of the SF ‘family’, an exception was made to the rule of only one instrument version. Both the SF-36 and the SF-12 were considered given that both have been used very widely in health research and neither represents a major burden to patients. The variety for this instrument family is complicated, with similar instruments also going by the names RAND-36 and RAND-12, and Veterans RAND 36 (VR-36) and VR-12. For the substantive parts of our work, the focus was only on SF-36 and SF-12.
Descriptive Overview of Short-listed PROMs (Project component 3)

For all short-listed PROMs, some basic descriptive information relating to the practical use of the instruments was seen as necessary to inform selection.

Methods

Information was sought from available literature, as well as the instrument websites and other online sources [28-59]. Instrument developers and authors were contacted for information not available in the literature or online sources [60,61].

For each short-listed instrument, the following information was sought.

Domains

For a review of the instrument domains, each instrument was applied to the PROMIS domain framework [24], which identifies seven subcomponent definitions (physical health symptoms, physical health function, mental health affect, mental health behaviour, mental health cognition, social health relationships, social health function). Each of the instruments was applied to the framework based on the number of items that coincided with each subcomponent category. The number of items, and the frequency of items falling within each domain category, was presented for comparison.

Response Options

Response options refer to the type of questions and scoring that the particular instrument applies for measurement. Common response options were identified and each instrument reviewed for use of each type.

- Likert scale (a statement which the respondent is asked to evaluate according to a set criteria; generally the level of agreement or disagreement is measured)
- Guttman scale (a set of survey items with binary answers that can be ranked in some order)
- Binary responses (e.g., Yes or No)
- Qualitative responses (e.g., free text responses)
- Other (including rating scales)

Burden on Respondent

Burden on respondent was identified as a combination of number of items, length of time for completion and word count. Each of these characteristics was identified and combined in a table for comparison across instruments.
Readability

The readability of the instrument was determined with the application of Microsoft’s Flesch Reading Ease test and the Flesch-Kincaid Grade Level test [25]. With the Reading Ease test, text is rated on a 100-point scale where a higher score indicates that the document is easier to understand. For most standard documents, a minimum score of 60 to 70 is desired [25]. The Flesch-Kincaid Grade Level test rates text on a US school grade level. Lower scores indicate a reading level accessible for readers with a lower grade level. For example, a reading score of 7.0 means that a seventh grader could be expected to understand the document. For increased readability, it is recommended that texts have a score of approximately 7.0 or 8.0 [25].

Recall Period

The recall period referred to the time frame each respondent is to consider when responding to questions presented in the instrument. The reported recall period was included for comparison.

Reported Modes and Methods of Administration

The method of administration of an instrument refers to the way in which the instrument is presented to the respondent (self-administered, interviewer, proxy, etc.). The mode of administration refers to the instrument format (in person, online, telephone, etc.). The reported available versions of each instrument were presented for comparison.

Translations

Translation refers to whether the instrument has been translated into the most common languages spoken in British Columbia: English, Chinese, Punjabi, Korean, Tagalog and Persian. In addition, the availability of a French version was also explored. Preferably the translated instrument would have undergone a full linguistic validation process. Sources for translation information included the PROQOLID website [26], the MAPI Research Trust website [23], and the NIH PROMIS website [27].

Instrument Cost

Information on the cost and restrictions on instrument use was obtained from instrument websites and/or instrument authors.

Results

Domains

Applying the PROMIS domain framework [24], Figure 2 shows that the mental and physical health domains were represented in all of the short-listed instruments. However, not all instruments included questions about social health and general health: the HUI3 does not cover
either of these two categories, and the AQoL-8D and NHP do not cover the general health domain.

![Figure 2. Domain coverage*](image)

Note: “domain coverage” refers to the representation of domains in the pool of items

**Response Options**

All the instruments with the exception of the NHP employed either Likert or Guttman-type response options, and several instruments included binary-type (e.g., yes-no) response options (see Table 2).

**Table 2. Response options**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Binary</th>
<th>Likert</th>
<th>Guttman</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL-8D</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SF-12®</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36®</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI3®</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>NHP</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QWB-SA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Burden on Respondent

Burden is an important characteristic in that it is linked to response: instruments that are less burdensome to respondents (e.g., not too long or complicated) are more likely to be completed. The EQ-5D, SF-12 and PROMIS/GHS have the fewest questions and take the least amount of time to complete (see Table 3).

**Table 3.** Respondent burden

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Number of Items</th>
<th>Word Count</th>
<th>Time for Completion (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL-8D</td>
<td>35</td>
<td>1,188</td>
<td>5</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>6</td>
<td>239</td>
<td>“few minutes”</td>
</tr>
<tr>
<td>SF-12®</td>
<td>12</td>
<td>453</td>
<td>2</td>
</tr>
<tr>
<td>SF-36®</td>
<td>36</td>
<td>692</td>
<td>10</td>
</tr>
<tr>
<td>HUI3®</td>
<td>15</td>
<td>1,173</td>
<td>8 - 10</td>
</tr>
<tr>
<td>NHP</td>
<td>38</td>
<td>353</td>
<td>5 - 15</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>80</td>
<td>1,934</td>
<td>15</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>26</td>
<td>607</td>
<td>5</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>10</td>
<td>217</td>
<td>2</td>
</tr>
</tbody>
</table>

Readability

The NHP, with its short sentences and simple words is the most readable instrument as measured by its reading ease score of 92.3 and the low grade level (grade 2) needed to understand the questions (see Table 4). In comparison, the EQ-5D requires a grade 11 reading level because it includes a very long question with multi-syllabic words. The AQoL-8D, SF-36, and QWB-SA are all similarly readable, requiring a grade level 5 to 6 and with their reading ease scores in the 70’s. The remaining instruments all fall within the recommended ranges for readability, with reading ease scores reaching a minimum of 60 and grade level scores below 8th grade [25].

**Table 4.** Readability

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Flesch Reading Ease</th>
<th>Flesch-Kincaide Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL-8D</td>
<td>73.0</td>
<td>5.3</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>68.8</td>
<td>10.6</td>
</tr>
<tr>
<td>SF-12®</td>
<td>72.2</td>
<td>7.1</td>
</tr>
</tbody>
</table>
Recall Period

The NHP, asking only questions in the current time context, has the shortest recall requirement (see Table 5). The SF-12, SF-36 and HUI3 include questions that are more difficult; asking for recall about events over the past 4 weeks.

Table 5. Recall period

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Current/ Now</th>
<th>Previous Week</th>
<th>Past 2 Weeks</th>
<th>Past 4 Weeks</th>
<th>Usual</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL-8D</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-12®</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SF-36®</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUI3®</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>NHP</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QWB-SA</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
</tbody>
</table>

Reported Modes and Methods of Administration

As Table 6 shows, the EQ-5D, SF-12 and SF-36 have the most administrative options, being tested for self-administration or interviewer and available in paper, telephone, interactive voice response for telephone, or online/electronic modes.
### Table 6. Administration methods and modes

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Methods</th>
<th>Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self</td>
<td>Interviewer</td>
</tr>
<tr>
<td>AQoL</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SF-12®</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SF-36®</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HUI3®</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NHP</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>QWB-SA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WHOQoL - BREF</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Note:** In Person/ Paper – provided to respondent in person via a paper based instrument; Telephone – available for use over the telephone; Online – available for use through an online source; Electronic – available for use on a computer; IVR - Interactive Voice Response for Telephone.

### Translations

The EQ-5D, SF-12 and SF-36 are available in most of the languages used frequently in British Columbia, whereas the AQoL and QWB-SA are only currently available in English and French (see Table 7) [23,26,27]. Some of the tools are available in several other languages (e.g., Dutch, Spanish, German, Greek, etc.). The translations listed in Table 7 may not have undergone a full linguistic validation process and may require further work to be suitable for use in a study.

### Table 7. Translations

<table>
<thead>
<tr>
<th>Instrument</th>
<th>English</th>
<th>Chinese</th>
<th>Punjabi</th>
<th>Korean</th>
<th>Tagalog</th>
<th>Farsi</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>SF-12®</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>SF-36®</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>HUI3®</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>NHP</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Instrument Cost

Many of the instruments are available for use for free: AQoL, NHP, WHOQoL-BREF and the PROMIS/GHS (see Table 8). And although the SF-36 is not free, the RAND and Veterans versions of the SF-36 are free for use (see Appendix 6 for additional information on instrument cost).

Table 8. Instrument cost

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>Free use</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Free use for research purposes only</td>
</tr>
<tr>
<td></td>
<td>Licensing fees depend on the type of study/trial/project, funding source, sample size and number of requested languages</td>
</tr>
<tr>
<td>SF-12®/SF-36®</td>
<td>Manual fees; licensing fees depend on the survey, the number of uses, the type of report requested</td>
</tr>
<tr>
<td></td>
<td>(*Note: Both the Rand and Veterans versions are free for use.)</td>
</tr>
<tr>
<td>HUI3®</td>
<td>Application and Interpretation Package costs $4000; additional questionnaires and manuals cost $2,000 each per study</td>
</tr>
<tr>
<td>NHP</td>
<td>Free use</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>Free use for research purposes only; otherwise $1000/yr plus $0.25 per use</td>
</tr>
<tr>
<td></td>
<td>Scoring: $57/hr or buy algorithm for scoring $240</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>Free use</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>Free use</td>
</tr>
</tbody>
</table>

(Appendix 6 reports additional information for each of the short-listed instruments [28-61].)
Review of Psychometric Properties of Short-listed PROMs (Project component 4a)

Given the project’s tight timeline and the potentially vast literature on psychometric evidence for all 9 instruments, a decision was made to search initially for review articles only. Therefore, the search was for published and grey literature review articles that discussed the psychometric properties of at least one of the 9 short-listed instruments.

Search Methods

Searches for literature reporting psychometric properties were developed in MEDLINE (OvidSP) and Embase (OvidSP) for the 9 short-listed instruments. None of the measures had MeSH subject headings in MEDLINE so keyword terms were used. The HUI was the only instrument in Embase that had an EMTREE subject heading. By only searching keyword terms there is the possibility that studies could be missed if the instrument was not mentioned in the title or abstract. A broader search was developed but this increased the numbers dramatically so it was not executed. COSMIN recommended a filter in PubMed for measurement properties which was reported in an article by Terwee et al. [62]. The filter used in the search had a sensitivity of 93.1 per cent and a precision of 9.4 per cent [62]. This filter was adapted for MEDLINE and Embase (OvidSP). A filter was also used for systematic reviews from the Scottish Intercollegiate Guidelines Network [63] as well as the publication type “reviews”. The search was comprised of three components; terms for the measures which were then combined with the measurement filter and then the review filter (see Appendix 4).

The MEDLINE and Embase search results were downloaded into RefWorks and placed in separate folders for each instrument. In total there were 1370 references. Duplicates were removed from each folder which left 1261 references in total. There were still duplicates as some papers were related to more than one instrument so they appeared in more than one folder. The folders contained the following number of references: AQoL (151), EQ-5D (200), HowsYourHealth (0), HUI (46), NHP (61), PROMIS (9), SF-36/12/RAND (717), QWB-SA (40) and WHOQoL-BREF (37). Separate grey literature searches were also performed resulting in 3 references for HowsYourHealth and 1 for PROMIS specifically focusing on the Global Health component of PROMIS. The two reviewers JCD and JB independently reviewed all of the references (see Appendix 5).

The majority of the instruments have websites that were also examined for evidence relating to psychometric properties.

Study Selection

The process of study selection is illustrated in Appendix 5. Our search yielded 1,261 articles, which discussed our candidate PROM instruments. The two reviewers JCD and JB
independently reviewed the titles and abstracts for each instrument and excluded those that focused solely on specific groups or disease conditions and that did not consider the psychometric properties of the instruments. JB selected 50 references and JCD selected 12. Discrepancies between the selections were reviewed by SB, who generated a final list of 21 articles for full text review (9 from JCD’s list and 21 from JB’s list). One other article that addresses the psychometric properties of the Global Health component of PROMIS was added to the list, therefore in total there were 22 articles reviewed [22,33,68-87]. There were no articles that address the psychometric properties of HowsYourHealth. The reviewers divided the list, and individually examined the text from each article for discussions on the internal consistency, reliability, validity, generalizability, responsiveness, and interpretability of the instruments.

**Data Extraction**

We developed a list of data extraction items for the studies included in this review. Each data extraction item includes a score, based on the COSMIN guidelines for that category [67]. The extracted information included the following: reliability (internal consistency and test-retest reliability), validity (content, construct, cross cultural and criterion validity), responsiveness, generalizability, and comparability with other candidate PROM instruments. No single article reported information on all extraction items. In cases where information was missing, but thought to be available, authors were contacted [60,61].

Two reviewers extracted data relevant to the psychometric properties of the instruments (JB and JCD). All text pertaining to the candidate PROM instruments was extracted by one of the reviewers and entered into a central database for subsequent data synthesis and quality assessment. As part of the data extraction, reviewers also recorded citation, psychometric property of interest, and PROM instrument being examined.

**Quality Assessment**

Because this review includes a wide variety of study designs and heterogeneity in study samples, we evaluated the quality of all studies to ensure that our coverage was representative and comprehensive, by adapting established protocols for evaluating systematic reviews and health instruments measurement properties [65,66]. Two reviewers independently evaluated all articles (JB and JCD) (see Appendix 8). Two reviewers also provided their overall assessment of the article, considering its breadth, scope, and presentation of findings on a 10 point scale. A comparison of these independent assessments indicated good coverage for all candidate instruments in most of the psychometric domains we intended to consider (see Results for further details).

In relation to our review of psychometric evidence, one limitation, noted in the quality assessment, was that the “review” search filter failed to exclude all non-review articles, due to a broader definition of the term “review” in bibliographic databases. While all the articles that we included were tagged as review articles in literature databases, our independent assessment
indicated that only 10 of these 22 articles were formal reviews of published literature. The remaining articles were either validation studies, or original research that compared how instruments performed within a specific sample.

**Data Synthesis**

We applied a scoring approach, proposed by the COSMIN-initiative, for systematic reviews of the measurement properties of instruments [67]. We considered aspects of each instrument’s reliability (internal consistency and test-retest reliability), validity (content, construct, cross cultural and criterion validity), and responsiveness, following strict guidelines on scoring in each of these categories [67]. We included two additional categories: generalizability and comparison with other PROMs instruments. For generalizability, we included any additional information about an instrument’s validity from evaluations of groups known to vary in their health status (e.g., in-patient vs. out-patient scores). We also recorded any strengths or weaknesses which were noted for using the instrument in any specific clinical group or population. Second, we recorded any text which directly compared two or more of our candidate instruments to one another. We did not provide scores in these categories, but instead provided a summary of relevant considerations. Two reviewers independently scored all extracted text (JB and JCD) and any discrepancies were reviewed by a third party (SB).

In applying our scoring criteria, we evaluated each article and all instruments discussed in that article individually, across 10 psychometric categories (listed below in Outcome Measures Section). Reviewers (JB and JCD) provided both a score for each category, and a text summary of their justification for that score based on the text extracted from each article. Scores ranged from very strong positive evidence (+++), to very strong negative evidence (---). Conflicting evidence was also noted (+/-), as was an absence of evidence in that particular category (?) (see Mokkink et al. [67] for further details on scoring criteria for each category). The COSMIN criteria were not directly applicable for all the forms of evidence pertaining to an instrument’s psychometric properties presented in the articles. In these instances, reviewers denoted that an alternative form of evidence was used for scoring, and provided the relevant text to justify their scores.

After evaluating all the text for each article, reviewers summarized the relevant text for each psychometric category, for each instrument, and provided a summary score. Summary scores and text were then compared. Any discrepancies were resolved through further discussion between the two reviewers (JB and JCD), and a final synthesized summary is provided in Appendix 7.

**Outcome Measures**

**Internal Consistency:** Internal consistency provides an estimate of the relatedness between these different questions and indicates if the assumptions about the relatedness between individual items, which were used to construct the scales and subscales, are valid [67]. To
receive a positive rating based on the COSMIN criteria in this category, instruments must have a Cronbach’s alpha(s) ≥ 0.70 and evidence that the (sub)scales are unidimensional.

**Reliability:** Reliability is most frequently assessed by examining the correlation between a patient’s score on a particular instrument at two different time points. Ideally, a patient’s scores should be identical if no important health change has occurred in the interim [67]. Large changes in scores may indicate that participants do not necessarily interpret the questions the same way at the different time points or responses are easily influenced by other factors. The COSMIN criteria for this category required Intraclass Correlation Coefficients (ICCs) or Kappas ≥ 0.70 or Pearson’s r ≥ 0.80 to receive a positive score.

**Content Validity:** Content validity is an assessment of both the comprehensiveness and relevance of an instrument for the population it is being used with. To truly assess content validity, one would need to examine these issues in relation to a particular population and research question. For example, some of the instruments focus primarily on physical health, while others include questions on both mental and physical health. Both types of instruments might have similar content validity for use in a research project focused on physical health, but vary in their content validity for research questions which also deal with issues pertaining to mental health. For the purpose of our review, we focused on assessments of content validity in a general population sample, and indicated where any particular limitations were noted. Positive scores based on the COSMIN criteria were assigned when evidence that the target population considered all items in the instrument to be relevant and considered the instrument complete.

**Construct Validity:** A variety of assumptions are frequently made in the construction of an instrument. For example, a developer may propose that an instrument measures two types of health (emotional and physical) and make decisions about which particular items are used to calculate scores on each of these aspects. An evaluation of an instrument’s construct validity provides an assessment of the degree to which these assumptions and hypotheses are valid based on data collected from participants [67]. The COSMIN criteria specified that factors that the instrument proposed to tap into needed to explain at least 50% of the variance to receive a positive score.

**Cross Cultural Validity:** A primary concern when an instrument is administered in multiple languages or in different cultural contexts is that differences in scores reflect true differences and not differences due to translation or alternative interpretations of questions. Strong cross cultural validity indicates that the instrument has been evaluated and performs similarly in multiple contexts [67]. Instruments needed to have their original factor structure confirmed, or studies which report no important differential item functioning, between different language versions or cultural contexts to receive a positive score according to the COSMIN criteria.

**Criterion Validity:** Criterion validity is assessed by evaluating the performance of an instrument in comparison to a “gold standard” instrument. The COSMIN criteria in this category require a convincing argument that there is a “gold standard” instrument and a correlation ≥ 0.70 between the “gold standard” and the candidate instrument, in order to achieve a positive rating.
**Responsiveness:** The ability of an instrument to detect change when it occurs in a patient is crucial if it is to be used for evaluating the efficacy of interventions. While information which allowed for scoring according to the COSMIN guidelines was not always reported, a great deal of other information on responsiveness was discussed in the reviews, including correlation with changes in clinical outcomes or groups known to vary in illness severity. The COSMIN criteria specify that an instrument should have a correlation ≥ 0.5 with another instrument measuring the same construct to receive a positive score.

**Generalizability (or External Validity):** We evaluated the extent to which each candidate PROM instrument could be generalized to other populations, and any examination of their ability to detect differences between groups known to vary in their health status.

**Comparison with candidate PROM instruments:** We recorded and evaluated all instances where articles directly compared the psychometric properties of our candidate PROM instruments.

**Results of Psychometric Review**

**Quality Assessment and Summary**

Twenty-one of the 22 articles provided information about the psychometric properties of the candidate PROM instruments (see Appendix 8) [22,33,68-87]. Ten of the articles also contained direct comparisons between two or more candidate instruments. All of the articles provided a good rationale and key objectives for the review, and most also provided a summary of their results in relation to the key objectives. Each of the 10 psychometric categories we considered was evaluated for most of our candidate instruments. However, it is important to note that none of the reviews explicitly discussed measurement error, an important psychometric category related to reliability.

An overview of the results from the psychometric review is provided in Table 9. A fuller summary of results is given in Appendix 7.

**Table 9. Overview of results from psychometric review**

<table>
<thead>
<tr>
<th></th>
<th>AQoL</th>
<th>EQ-5D</th>
<th>SF-36</th>
<th>HUI</th>
<th>NHP</th>
<th>QWB-SA</th>
<th>WHOQoL-BREF</th>
<th>PROMIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency</td>
<td>+</td>
<td>n/a</td>
<td>+++</td>
<td>?</td>
<td>+/-</td>
<td>n/a</td>
<td>+/-</td>
<td>+++</td>
</tr>
<tr>
<td>Reliability</td>
<td>?</td>
<td>+/-</td>
<td>+++</td>
<td>+/-</td>
<td>+/-</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Content validity</td>
<td>-</td>
<td>-</td>
<td>++</td>
<td>+/-</td>
<td>+/-</td>
<td>+/-</td>
<td>?</td>
<td>+/-</td>
</tr>
<tr>
<td>Construct validity</td>
<td>?</td>
<td>?</td>
<td>+++</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Cross-cultural validity</td>
<td>?</td>
<td>+</td>
<td>+/-</td>
<td>+</td>
<td>+/-</td>
<td>?</td>
<td>+++</td>
<td>?</td>
</tr>
</tbody>
</table>

28
Internal Consistency

The majority of our candidate PROM instruments are reflective measures, utilizing several questions to provide an overall estimate for a domain or subdomain of a participant’s health state. There was a large evidence base which indicated that the SF-36 performed particularly well in this category for its scales and subscales [33,68-71]. PROMIS also performed well, but the evidence base was smaller [72]. The AQoL was reported as having good internal consistency [73], but no numerical estimates which allowed assessment compared to the COSMIN guidelines were provided in our sample of review papers. Both the WHOQoL-BREF and the NHP instruments did not meet the COSMIN criteria for at least one of the subscales, which resulted in relatively lower scores on this psychometric aspect [33,74,75]. Internal consistency was not considered for the EQ-5D and QWB-SA, as they have single questions for each domain or subdomain.

Reliability

The SF-36 had the most extensive and strongest evidence for its reliability on its two main subscales [69,70]. The EQ-5D, HUI and NHP had conflicting evidence, and reported correlations between scores were both above and below the COSMIN thresholds in this domain [22,33,70,76,77]. For the remaining instruments, test-retest reliability was not reported (PROMIS, WHOQoL-BREF and AQoL), or methodological issues were raised with existing assessments (QWB-SA) [22].

Content Validity

Both the SF-36 and PROMIS were noted to be fairly comprehensive and relevant for use in a general population sample [33,71,72]. Both the HUI and QWB-SA were noted to have limitations in their coverage of mental health, but were otherwise cited as being comprehensive [22,33]. The NHP is stated to have good content validity for working with populations with major burdens of disease, but may be less relevant in a general population sample [33,74] . The authors of one review noted the failure to find evidence of the AQoL’s content validity, despite extensive searching within the published literature [73]. Content validity was not discussed for the WHOQoL-BREF in any of our review articles.
Construct Validity

From the evidence presented in our reviews, the SF-36, the WHOQoL-BREF, and PROMIS all had positive evidence for their construct validity [68,69,71,72,75]. The NHP was the only instrument where evidence was reported which did not meet the COSMIN standards for construct validity [76]. No evidence was reported in the reviews articles for the AQoL, EQ-5D, QWB-SA, and HUI, which allowed us to assess their content validity.

Cross Cultural Validity

The WHOQoL-BREF, which prioritized cross-cultural validity during its initial development, performed best in this category. This instrument has been systematically evaluated in many different contexts, and only a few meaningful differences have been noted [75]. The EQ-5D and HUI both are used in different cultural and linguistic contexts, providing some positive evidence, but nothing pertaining to the COSMIN criteria was reported in our reviews. The SF-36 is also widely used in many contexts, but some evidence of validation across different groups (using differential item functioning analysis) has been reported for both the SF-36 and SF-12 [69,79]. The NHP had positive evidence in terms of how it was developed and its widespread use [33,74], but one review noted potential differences between how the instrument functions in French versus English [74]. There was no information reported in our reviews that allowed us to assess the cross cultural validity of the QWB-SA, PROMIS and the AQoL.

Criterion Validity

As there is no widely accepted “gold standard” PROM instrument [22,33,73,76], scoring in this domain is challenging. However, correlations between our candidate instruments, reported in Appendix 7, do provide us with some useful information about criterion validity. The SF-36 and the EQ-5D were the most common instruments utilized for comparison in the absence of a gold standard, adding some support for their own validity. PROMIS met the COSMIN criteria based on correlations with the EQ-5D [72]. Although the EQ-5D did meet the same criteria when its correlation with the SF-12 physical component summary score was examined, it did not meet the criteria for its correlation with the mental component summary score [33].

Responsiveness

The SF-36 was the clear leader in this category, with multiple papers citing not only its strong performance in this domain, but also its greater responsiveness compared to other candidate instruments [71,72,76,77,80]. PROMIS also scored well in this category, but the evidence base was substantially smaller than for the SF-36 [72]. The AQoL also received a positive score, based on its correlation with groups known to vary in illness severity, but its evidence base was also limited [73]. While the QWB-SA met the COSMIN criteria in this category, issues were raised with insensitivity of the questions to detect changes in the frequency or intensity of conditions due to the dichotomous nature of the questions [22]. The HUI also had conflicting evidence. While it was cited to be able to detect differences between known groups [81], over 20% of a general population received perfect scores, indicating an insensitivity to detect more
minor health issues [80]. While the NHP has been shown to be sensitive in samples with major health issues, and possibly even more sensitive than the SF-36 [70], its inability to detect change in more healthy individuals is an issue for working with a general population [82]. The EQ-5D is noted to not be very responsive compared to our other candidate instruments [22,33,74,80], although this may partially be attributed to its substantially shorter length.

**Generalizability and Head to Head Comparisons**

The SF-36 seemed to perform particularly well compared to other instruments, especially in the category of responsiveness [33,70,76,77,80]. Most instruments produced scores which reflected differences between groups of patients known to vary in their illness severity, further speaking to their validity for measuring patient’s health. Several issues that did not directly fall into the other psychometric categories were noted here as well. The NHP was noted to be particularly useful when working with populations with major burdens of disease, but less useful in a general population setting [33,74,82]. While most of the candidate instruments are widely used, one review did make note that the majority of research using the HUI has been conducted by one group of researchers at McMaster University [22]. Finally, while the QWB-SA is more likely to capture additional variation in a primarily healthy population, this may in part be due to overweighting of minor conditions (e.g. wearing eyeglasses) [22].
Decision-making Criteria (Project component 4b)

For the short-listed instruments, the formal review on psychometric properties was supplemented with information relating to selected decision-making criteria. These criteria were developed to consider instrument appropriateness for policy, management and clinical practice. The criteria included:

- Norm reference sets (to allow comparison of sample data to the general population);
- Utility/preference scoring algorithm (to all the calculation of quality-adjusted life years and so facilitate cost-utility analyses);
- Clinically relevant thresholds (to inform judgments on the value of observed changes in scores).

Information on these criteria was obtained for each instrument through examination of the appropriate website or scoring guides. In some cases, instrument sources identified literature validating the use of scoring methods and outcomes, while other instrument sources simply outlined their reasoning for scoring and outcome structures. Wherever possible, reported normative data were included as well as scoring algorithms. Total scores or global scores, score types (i.e. dimension, item scores, utility scores, non-reference, other) were identified. The reported scoring methodology and outcome was included for each instrument.

Although the AQoL, SF-12, SF-36 and QWB-SA all supply utility scores, Table 10 shows that only EQ-5D and HUI3 provide utility scores derived from a Canadian sample. Likewise, all the instruments provide population norms, but only the EQ-5D, SF-12/SF-36 and HUI3 supply Canadian population norms. The majority of instruments (bar the NHP, WHOQoL-BREF and PROMIS-GHS) publish minimally important differences, which provide an indication of the size of change in score required for it to be viewed as a clinically important change.

**Table 10. Utility scores, population norms, minimally important differences**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Utility Scores</th>
<th>Population Norms</th>
<th>Minimally Important Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>✓</td>
<td>✓(CDN)</td>
<td>✓</td>
</tr>
<tr>
<td>SF-12®</td>
<td>✓</td>
<td>✓(CDN)</td>
<td>✓</td>
</tr>
<tr>
<td>SF-36®</td>
<td>✓</td>
<td>✓(CDN)</td>
<td>✓</td>
</tr>
<tr>
<td>HUI3®</td>
<td>✓(CDN)</td>
<td>✓(CDN)</td>
<td>✓</td>
</tr>
<tr>
<td>NHP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QWB-SA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument</td>
<td>Utility Scores</td>
<td>Population Norms</td>
<td>Minimally Important Differences</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: CDN indicates Canadian-specific utility or population norms available.
Additional Information (Project component 5)

For the short-listed PROM instruments, additional information was identified, including examples of use in a primary and community care context (with a focus on primary care reform or practice change) and PROM-related activity in other jurisdictions (with a particular focus on other Canadian jurisdictions but also the UK and US).

In order to ensure the selected instruments had been used in a primary and community care context a search was conducted through the Patient Reported Outcome Measurement Group (date limit for search ended at 2005) and the MEDLINE and Embase databases. The PROM Bibliography on the Patient Reported Outcome Measurement Group’s website has the option to limit searches to specific instruments and set keywords. All instruments were selected and then combined with primary care. Searches in MEDLINE and Embase focused on the instruments being combined with key primary health terms (primary health care, public health, Health Promotion, Primary Prevention, Preventive Health Services, Preventive Medicine). SB reviewed the titles and abstracts from the Patient Reported Outcome Measurement Group’s PROM Bibliography search and a select number from the MEDLINE and Embase searches.

Table 11 gives an indication of the quantity of studies undertaken in a primary care setting using each of the selected instruments. Searching using the PROMIS acronym again proved difficult and so results are not reported for that instrument. Clearly both the SF instruments and EQ-5D stand out as being the most widely applied in primary and community care contexts.

Table 11. Use of instruments in primary care settings

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Patient Reported Outcome Measurement Group</th>
<th>MEDLINE Expanded Search</th>
<th>Embase Expanded Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>2</td>
<td>49</td>
<td>73</td>
</tr>
<tr>
<td>SF-12/ SF-36®</td>
<td>41</td>
<td>239</td>
<td>307</td>
</tr>
<tr>
<td>HUI3®</td>
<td>2</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>NHP</td>
<td>4</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>QBW-SA</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>0</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

From a variety of sources available to the project team, intelligence was gathered on PROM-related activity in other jurisdictions including other Canadian provinces and territories, the US and the UK. Table 12 below gives an indication of some of the work identified.
**Table 12. Examples of use of instruments in other jurisdictions**

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Organisation(s)</th>
<th>Project / Area</th>
<th>PROM(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberta</td>
<td>Alberta Health Services</td>
<td>Alberta's Caring for Diabetes Project</td>
<td>EQ-5D</td>
</tr>
<tr>
<td>Alberta</td>
<td>Alberta Health Services</td>
<td>Province-wide use of PROMs</td>
<td>EQ-5D, SF-12</td>
</tr>
<tr>
<td>Alberta</td>
<td>University of Alberta Hospital, Edmonton</td>
<td>Health &amp; Lung Transplant</td>
<td>HUI</td>
</tr>
<tr>
<td>Alberta</td>
<td>Alberta Bone &amp; Joint Health Institute</td>
<td>Hip / knee replacement</td>
<td>SF-36</td>
</tr>
<tr>
<td>BC</td>
<td>All regional health authorities</td>
<td>Knee arthroplasty (PEAK project)</td>
<td>SF-12, EQ-5D</td>
</tr>
<tr>
<td>BC</td>
<td>Vancouver Coastal Health</td>
<td>Elective surgery (VALHUE project)</td>
<td>EQ-5D</td>
</tr>
<tr>
<td>Manitoba</td>
<td>Winnipeg</td>
<td>Hip / knee replacement</td>
<td>SF-12</td>
</tr>
<tr>
<td>Ontario</td>
<td>Cancer Care Ontario</td>
<td>Ontario PROs of Symptoms and Toxicity</td>
<td>EQ-5D</td>
</tr>
<tr>
<td>Canada</td>
<td>Statistics Canada</td>
<td>Canadian Community Health Survey</td>
<td>HUI, RAND</td>
</tr>
<tr>
<td>England</td>
<td>NHS England</td>
<td>Selected elective surgeries</td>
<td>EQ-5D</td>
</tr>
<tr>
<td>UK</td>
<td>BUPA/Spire Healthcare (private sector)</td>
<td>Elective surgical procedures</td>
<td>SF-36</td>
</tr>
<tr>
<td>US</td>
<td>Medicare</td>
<td>Health Outcomes Survey</td>
<td>VR-12</td>
</tr>
<tr>
<td>US</td>
<td>Veterans Administration</td>
<td>Various studies</td>
<td>VR-36, VR-12</td>
</tr>
</tbody>
</table>
Workshop and Recommendations

The project culminated with a workshop, held at the BC Ministry of Health on January 14, 2013, involving a wide range of stakeholders. The participants included the research team, research colleagues from the Ministry of Health and the Michael Smith Foundation for Health Research, representatives from all BC health authorities, one colleague from CIHR, and others from various parts of the BC health system.

The objectives for the whole workshop were:

- To provide an understanding of the basics of PROMs
- To update on the Integrated Primary and Community Care (IPCC) policy context for the use of PROMs in BC
- To share the details of the review and evidence synthesis work undertaken by the research team
- To have reflection and deliberation on the PROM instruments and their potential use in IPCC projects
- To come to a consensus on which PROM instrument(s) should be recommended for use in BC

The main output of the workshop, therefore, was seen as being a recommendation on instrument selection.

In broad terms there was consensus that the following instruments performed well on the chosen criteria:

- Domain coverage: SF, EQ-5D, QWB-SA, WHOQoL-BREF, PROMIS
- Burden and readability: SF, PROMIS, AQoL
- Cost: AQoL, NHP, WHOQoL-BREF, PROMIS
- Administration options and translations: EQ-5D, SF
- Utility scale and population norms: EQ-5D, HUI, SF
- Psychometrics:
  - Internal consistency: SF, PROMIS
  - Reliability: SF
  - Content validity: SF
  - Construct validity: WHOQoL-BREF, SF
  - Cross cultural validity: WHOQoL-BREF
  - Responsiveness: SF

There were two rounds of voting and discussion at the Workshop which occurred after information was presented on the review and participants had an opportunity to visually review the items within each measure. The final votes indicated a strong preference for two instruments, PROMIS (41 votes) and SF (36 votes), and a desire to keep EQ-5D (18 votes) as a back-up.
An evaluation of the workshop was undertaken at the end and the results are reported as Appendix 10.
Appendix 1. Cited References in the Web of Science

Each measure will usually have more than one paper indexed in the Web of Science. Total Cites* indicates the total number of cites from the papers searched that pertained to a measure.

Table 13. Cited references in the Web of Science

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full name</th>
<th>Author(s)</th>
<th>Publication Year</th>
<th>Total Cites*</th>
</tr>
</thead>
<tbody>
<tr>
<td>howRU™</td>
<td>howRU™</td>
<td>Benson Tim Whatling Justin</td>
<td>2010</td>
<td>1</td>
</tr>
<tr>
<td>HSQ</td>
<td>Health Status Questionnaire 2.0</td>
<td>Radosevich David M</td>
<td>1999</td>
<td>10</td>
</tr>
<tr>
<td>TAAQOL</td>
<td>TNO-AZL Questionnaire for Adult's Health-related Quality of Life</td>
<td>Bruil Jeanet Fekkes Minne Verrips Erik GH et al.</td>
<td>2001</td>
<td>25</td>
</tr>
<tr>
<td>QLQ-E</td>
<td>Quality of Life Questionnaire-Evans</td>
<td>Evans Cope Wendy Evans David</td>
<td>1985</td>
<td>26</td>
</tr>
<tr>
<td>SQLP</td>
<td>Subjective Quality of Life Profile</td>
<td>Dazord Alice</td>
<td>1991</td>
<td>37</td>
</tr>
<tr>
<td>GQOL</td>
<td>Global Quality of Life Scale</td>
<td>Hyland Michael E Sodergren Samantha C</td>
<td>1996</td>
<td>43</td>
</tr>
<tr>
<td>HALex</td>
<td>Health and Activity Limitation Index</td>
<td>Bradley Cathy J Erikson Pennifer</td>
<td>1998</td>
<td>55</td>
</tr>
<tr>
<td>HowsYourHealth</td>
<td>HowsYourHealth</td>
<td>Trustees of Dartmouth College &amp; FNX Corp.</td>
<td>1992</td>
<td>163</td>
</tr>
<tr>
<td>DUKE</td>
<td>Duke Health Profile</td>
<td>Parkerson George R</td>
<td>1990</td>
<td>168</td>
</tr>
<tr>
<td>MHIQ</td>
<td>McMaster Health Index Questionnaire</td>
<td>Chambers Larry W</td>
<td>1982</td>
<td>173</td>
</tr>
<tr>
<td>PGI</td>
<td>Patient Generated Index</td>
<td>Garratt Andrew M Ruta Danny A</td>
<td>1994</td>
<td>340</td>
</tr>
<tr>
<td>QOLS</td>
<td>Flanagan's Quality of Life Scale</td>
<td>Flanagan John C</td>
<td>1978</td>
<td>402</td>
</tr>
<tr>
<td>15D©</td>
<td>15-dimensional health-related quality of life measure</td>
<td>Sintonen Harri</td>
<td>2001</td>
<td>407</td>
</tr>
<tr>
<td>QOLI®</td>
<td>Quality of Life Inventory®</td>
<td>Frisch Michael B</td>
<td>1992</td>
<td>412</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full name</td>
<td>Author(s)</td>
<td>Publication Year</td>
<td>Total Cites*</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>COOP-C or COOP/WONCA</td>
<td>COOP/WONCA Charts</td>
<td>Nelson Eugene C Wasson John H</td>
<td>1987</td>
<td>499</td>
</tr>
<tr>
<td>AQoL</td>
<td>Assessment of Quality of Life</td>
<td>Hawthorne Graeme Osborne Richard H Richardson Jeff</td>
<td>1999</td>
<td>538</td>
</tr>
<tr>
<td>SEIQoL</td>
<td>Schedule for the Evaluation of Individual Quality of Life</td>
<td>Browne J Hickey Anne M Hiltbrunner B et al.</td>
<td>1994</td>
<td>699</td>
</tr>
<tr>
<td>PROMIS</td>
<td>Patient-Reported Outcomes Measurement Information System</td>
<td>NIH</td>
<td>1973?</td>
<td>1032</td>
</tr>
<tr>
<td>QWB</td>
<td>Quality of Well Being scale</td>
<td>Anderson John Kaplan Robert M</td>
<td>1989</td>
<td>1156</td>
</tr>
<tr>
<td>IPQ-R</td>
<td>Revised Illness Perception Questionnaire</td>
<td>Moss-Morris Rona Buick Deanna Cameron Linda D et al.</td>
<td>2002</td>
<td>1321</td>
</tr>
<tr>
<td>RAND-36</td>
<td>The RAND 36-Item Health Survey 1.0</td>
<td>RAND</td>
<td></td>
<td>1403</td>
</tr>
<tr>
<td>NHP</td>
<td>Nottingham Health Profile</td>
<td>Hunt S McEwen J McKenna Stephen P</td>
<td>1981</td>
<td>3154</td>
</tr>
<tr>
<td>HUI®</td>
<td>Health Utilities Index</td>
<td>Feeny David Furlong W Torrance George W</td>
<td>1995</td>
<td>4225</td>
</tr>
<tr>
<td>WHOQoL-100 &amp; WHOQoL-BREF</td>
<td>World Health Organization Quality of Life assessment instrument</td>
<td>The WHOQoL group</td>
<td>1993</td>
<td>6294</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Euroqol EQ-5D</td>
<td>The EuroQol Group</td>
<td>1990</td>
<td>6372</td>
</tr>
<tr>
<td>SF-12® / SF-12v2™</td>
<td>SF-12® Health Survey and SF-12v2™ Health Survey</td>
<td>Ware John E</td>
<td>1996</td>
<td>6512</td>
</tr>
<tr>
<td>SIP</td>
<td>Sickness Impact Profile</td>
<td>Bergner Marilyn Gilson Betty</td>
<td>1981</td>
<td>7748</td>
</tr>
<tr>
<td>SF-36® / SF-36v2™</td>
<td>SF-36® Health Survey and SF-36v2™ Health Survey</td>
<td>Sherbourne Cathy D Ware John E</td>
<td>1992</td>
<td>13874+</td>
</tr>
</tbody>
</table>
Appendix 2. Patient Reported Outcomes Measurement Group Bibliography

Latest entry in the Patient Reported Outcomes Measurement Group database is 2005. A subject and keyword search were completed. T1 was a keyword search of author list, title or abstract. T2 search was using the defined list of instrument names. There may be overlap between the two searches.

Table 14. Patient reported outcomes measurement group bibliography search

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full name</th>
<th>T1</th>
<th>T2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAAQOL (2001)</td>
<td>TNO-AZL Questionnaire for Adult’s Health-related Quality of Life</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>QLQ-E (1985)</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>howRU™ (2010)</td>
<td>howRU™</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>HowsYourHealth (1992)</td>
<td>HowsYourHealth</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>PROMIS (1973?)</td>
<td>Patient Reported Outcomes Measurement Information System</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>RAND-36 (1993?)</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>IPQ-R</td>
<td>Revised Illness Perception Questionnaire</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GQOL</td>
<td>Global Quality of Life Scale</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>HAlex</td>
<td>Health and Activity Limitation Index</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>QWB-SA</td>
<td>Quality of Well-Being scale Self-Administered</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>MHIQ</td>
<td>McMaster Health Index Questionnaire</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>QOLS</td>
<td>Flanagan’s Quality of Life Scale</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>SQLP</td>
<td>Subjective Quality of Life Profile</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>15D©</td>
<td>15-dimensional health-related quality of life measure</td>
<td>28</td>
<td>11</td>
</tr>
<tr>
<td>PGI</td>
<td>Patient Generated Index</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>DUKE</td>
<td>Duke Health Profile</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>COOP-C or COOP/WONCA</td>
<td>COOP/WONCA Charts</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>SEIQoL</td>
<td>Schedule for the Evaluation of Individual Quality of Life</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>HSQ</td>
<td>Health Status Questionnaire 2.0</td>
<td>99</td>
<td>13</td>
</tr>
<tr>
<td>QWB</td>
<td>Quality of Well Being scale</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full name</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------------------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>QOLI®</td>
<td>Quality of Life Inventory®</td>
<td>130</td>
<td>22</td>
</tr>
<tr>
<td>HUI®</td>
<td>Health Utilities Index</td>
<td>189</td>
<td>178</td>
</tr>
<tr>
<td>WHOQoL-100 &amp; WHOQoL-BREF</td>
<td>World Health Organization Quality of Life assessment instrument</td>
<td>200</td>
<td>7</td>
</tr>
<tr>
<td>AQoL</td>
<td>Assessment of Quality of Life</td>
<td>218</td>
<td>23</td>
</tr>
<tr>
<td>NHP</td>
<td>Nottingham Health Profile</td>
<td>304</td>
<td>288</td>
</tr>
<tr>
<td>SF-12® / SF-12v2™</td>
<td>SF-12® Health Survey and SF-12v2™ Health Survey</td>
<td>335</td>
<td></td>
</tr>
<tr>
<td>SIP</td>
<td>Sickness Impact Profile</td>
<td>345</td>
<td></td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Euroqol EQ-5D</td>
<td>469</td>
<td>428</td>
</tr>
<tr>
<td>SF-36® / SF-36v2™</td>
<td>SF-36® Health Survey and SF-36v2™ Health Survey</td>
<td>2367</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 3. Total Citations for the Last Six Years

The Web of Science records for the cited references of the 28 measures were reviewed. The number of cites for the last 6 years were recorded and tabulated and then sorted by the total number of cites for this time period.

- Each number in the year column represents the total number of times all the searched articles on a measure were cited
- The totals have been aggregated for all years, the last 6 years and then separately for each year. The years 2011/2012 are combined as there were very few for 2012.

Table 15. Total citations for the years 2007 to 2012

<table>
<thead>
<tr>
<th></th>
<th>PROMS Instrument</th>
<th>Publication Year</th>
<th>Total Cites All Years</th>
<th>Total Cites Last 6 Years</th>
<th>Cites for 2011/2012</th>
<th>Cites for 2010</th>
<th>Cites for 2009</th>
<th>Cites for 2008</th>
<th>Cites for 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EQ-5D: Euroqol EQ-5D</td>
<td>1990</td>
<td>6372</td>
<td>1599</td>
<td>63</td>
<td>243</td>
<td>431</td>
<td>286</td>
<td>556</td>
</tr>
<tr>
<td></td>
<td>SF-36® Health Survey and SF-36v2™ Health Survey</td>
<td>1992</td>
<td>13874+</td>
<td>1241</td>
<td>39</td>
<td>75</td>
<td>264</td>
<td>490</td>
<td>373</td>
</tr>
<tr>
<td></td>
<td>SF-12® Health Survey and SF-12v2™ Health Survey</td>
<td>1996</td>
<td>6512</td>
<td>301</td>
<td>4</td>
<td>34</td>
<td>52</td>
<td>112</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Rand-36 The RAND 36-Item Health Survey 1.0</td>
<td>1993</td>
<td>1403</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>PROMIS: Patient-Reported Outcomes Measurement Information System</td>
<td>2007</td>
<td>1032</td>
<td>982</td>
<td>26</td>
<td>129</td>
<td>91</td>
<td>166</td>
<td>570</td>
</tr>
<tr>
<td>3</td>
<td>WHOQoL-100 &amp; WHOQoL-BREF</td>
<td>1993</td>
<td>6294</td>
<td>295</td>
<td>26</td>
<td>37</td>
<td>73</td>
<td>92</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>HUI Health Utilities Index</td>
<td>1995</td>
<td>4225</td>
<td>135</td>
<td>7</td>
<td>17</td>
<td>64</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>PROMS Instrument</td>
<td>Publication Year</td>
<td>Total Cites All Years</td>
<td>Total Cites Last 6 Years</td>
<td>Cites for 2011/2012</td>
<td>Cites for 2010</td>
<td>Cites for 2009</td>
<td>Cites for 2008</td>
<td>Cites for 2007</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------</td>
<td>------------------</td>
<td>-----------------------</td>
<td>--------------------------</td>
<td>----------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>6</td>
<td>IPQ-R: Revised Illness Perception Questionnaire</td>
<td>2002</td>
<td>1321</td>
<td>33</td>
<td>4</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>• NHP: Nottingham Health Profile</td>
<td>1981</td>
<td>3154</td>
<td>30</td>
<td>1</td>
<td></td>
<td>13</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sickness Impact Profile</td>
<td>1981</td>
<td>7748</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>QWB-SA: Quality of Well Being scale-Self Administered</td>
<td>1998</td>
<td>174</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>HowsYourHealth</td>
<td>1992</td>
<td>163</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>AQoL: Assessment of Quality of Life</td>
<td>1999</td>
<td>538</td>
<td>17</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 4. Search Strategy for Reviews of PROMS Instruments’ Psychometric Properties

Database: Ovid MEDLINE(R) 1946 to Present with Daily Update

Search Strategy:

-----------------------------------------------

General Search for the 9 Selected Measures

1     SF-36$.tw. (10434)
2     SF36$.tw. (602)
3     Short Form 36$.tw. (5019)
4     Short-Form36$.tw. (1)
5     RAND36$.tw. (6)
6     RAND 36$.tw. (430)
7     SF 12$.tw. (1679)
8     SF12$.tw. (144)
9     Short-Form 12$.tw. (658)
10    Short-Form12$.tw. (2)
11    or/1-10 (15371)

12     EQ 5D$.tw. (1932)
13     Euroqol$.tw. (1603)
14     EQ5D$.tw. (105)
15     or/12-14 (2793)

16     patient reported outcomes measurement information system.tw. (72)
17     PROMIS.tw. (148)
18     Information Systems/ (17640)
19     "National Institutes of Health (U.S.)"/ (14957)
20     "Outcome Assessment (Health Care) "/ (43061)
21     and/18-20 (11)
22     or/16-17,21 (158)

23     World Health Organization Quality of Life assessment instrument.tw. (29)
24     WHOQOL$.tw. (893)
WHO Quality of Life Assessment Instrument.tw. (5)
World Health Organization Quality of Life Assessment.tw. (72)
or/23-26 (920)

(Health Utilities Index or Health Utility Index).tw. (522)
HUI-2.tw. (23)
HUI2.tw. (101)
HUI 3.tw. (43)
HUI3.tw. (180)
or/28-32 (553)

nottingham health profile$.tw. (926)
sickness impact profile/ (5488)
sickness impact profile?.tw. (954)
or/34-36 (6959)

quality of well being scale?.tw. (144)
QWB$.tw. (178)
or/38-39 (250)

"HowsYourHealth".tw. (2)
hows your health.tw. (2)
or/41-42 (2)

Assessment of quality of life.tw. (994)
AQoL-4D.tw. (0)
AQoL-8D.tw. (0)
AQoL-7D.tw. (1)
AQoL-8D.tw. (0)
AQoL$.tw. (69)
or/44-49 (1022)

or/11,15,22,27,33-34,40,43,50 (22532) All Measures Combined

COSMIN Measurement Filter (Lines 52-101)
instrumentation.tw. (20143)
Validation Studies/ (57848)
"reproducibility of results"/ (238617)
reproducib$.tw. or psychometrics/ or psychometr$.tw. or clinimetr$.tw. or clinometr$.tw. or observer variation/ or observer variation.tw. or discriminant analysis/ or reliab$.tw. or valid$.tw. or coefficient.tw. or internal consistency.tw. (710911)
(cronbach$ and (alpha or alphas)).tw. (7954)
(item correlation or item correlations).tw. (287)
(item selection or item selections or item reduction).tw. (544)
(item reductions or agreement or precision).tw. (184459)
(imprecision or precise values).tw. (3384)
test-retest.tw. (12043)
test and retest).tw. (12765)
(reliab$ and (test or retest)).tw. (45671)
(stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intra-tester or intra-tester).tw. (195750)
(interobserver or inter-observer or intraobserver or intra-observer).tw. (14015)
(intertechnician or intertechnician or intratechnician or intra-technician).tw. (7)
(interexaminer or inter-examiner or intraexaminer or intra-examiner).tw. (1172)
(interassay or inter-assay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-individual or interparticipant or inter-participant or intraparticipant or intra-participant).tw. (27244)
(kappa or kappa s or kappas).tw. (48660)
(coefficient of variation or repeatab$).tw. (28500)
((replicab$ or repeated) and (measure or measures or findings or result or results or test or tests)).tw. (112067)
(generaliza$ or generalisa$ or concordance).tw. (41275)
(intraclass and correlation$).tw. (9257)
(discriminative or known group or factor analysis or factor analyses or factor structure or factor structures).tw. (32272)
(dimensionality or subscale$ or multitrait scaling analysis or multitrait scaling analyses or item discriminant or interscale correlation or interscale correlations).tw. (22862)
((error or errors) and (measure$ or correlat$ or evaluat$ or accuracy or accurate or precision or mean)).tw. (87795)
(individual variability or interval variability or rate variability or variability analysis).tw. (14119)
(uncertainty and (measurement or measuring)).tw. (2680)
(standard error of measurement or sensitiv$ or responsive$).tw. (954073)
(limit and detection).tw. (40318)
(minimal detectable concentration or interpretab$$.tw. (4261)
(small$ and (real or detectable) and (change or difference)).tw. (3625)
(meaningful change or minimal important change or minimal important difference or minimally important change or minimally important difference or minimal detectable change or minimal detectable difference or minimally detectable change or minimally detectable difference or minimal real change or minimal real difference or minimally real change or minimally real difference).tw. (804)
(ceiling effect or floor effect or Item response model or IRT or Rasch or Differential item functioning or DIF or computer adaptive testing or item bank or cross-cultural equivalence).tw. (5189)
or/52-84 (2203306)

exp addresses/ (9536)
biography/ or comment/ or editorial/ or letter/ or news/ (1434252)
case reports/ (1589869)
directory/ (6620)
festschrift/ (1353)
interview/ (22507)
lectures/ (5570)
legal cases/ or legislation/ (11581)
newspaper article/ or exp popular works/ (21277)
patient education handout/ (3574)
exp congresses/ (60630)
consensus development conference/ or consensus development conference, nih/ (8343)
practice guideline/ (17065)
or/86-98 (3003509)
animals/ not humans/ (3680958)
99 not 100 (2919500)
Selected Measures AND the Measurement Filter

102     (11 and 85) not 101 [SF-36 & SF-12] (6200)
103     (15 and 85) not 101 [EQ-5D] (1153)
104     (22 and 85) not 101 [PROMIS] (102)
105     (27 and 85) not 101 [WHOQOL] (430)
106     (33 and 85) not 101 [HUI] (298)
107     (34 and 85) not 101 [Nottingham Health Profile] (367)
108     (40 and 85) not 101 [QWB] (146)
109     (43 and 85) not 101 [Hows Your Health] (0)
110     (50 and 85) not 101 [AQoL] (421)
111     (51 and 85) not 101 [All Measures] (9028)
112     limit 111 to yr="1990 -Current" (8986)
113     limit 112 to English language (8396)

114     (51 and 85) not 101 [All Measures] (9028)
115     limit 114 to "review articles" (362)

Systematic Review Filter for MEDLINE

116     Meta-Analysis as Topic/ (12379)
117     meta analy$.tw. (41533)
118     metaanaly$.tw. (1093)
119     Meta-Analysis/ (35816)
120     (systematic adj (review$1 or overview$1)).tw. (33411)
121     exp Review Literature as Topic/ (6466)
122     or/116-121 (85325)
123     cochrane.ab. (20389)
124     embase.ab. (17972)
125     (psychlit or psyclit).ab. (843)
126     (psychinfo or psycinfo).ab. (6377)
127     (cinahl or cinhal).ab. (6822)
128     science citation index.ab. (1527)
129     bids.ab. (312)
130     cancerlit.ab. (536)
131     or/123-130 (32642)
132     reference list$.ab. (7310)
133     bibliograph$.ab. (9635)
134     hand-search$.ab. (3033)
relevant journals.ab. (541)
manual search$.ab. (1752)
or/132-136 (19928)
selection criteria.ab. (16126)
data extraction.ab. (7414)
138 or 139 (22261)
Review/ (1729654)
140 and 141 (15658)
Comment/ (484242)
Letter/ (757777)
Editorial/ (305167)
animal/ (5031848)
human/ (12508105)
146 not (146 and 147) (3680958)
or/143-145,148 (4784933)
150 or 131 or 137 or 142 (108703)
151 150 not 149 [Systematic Review Filter] (101078)

Selected Measures AND Measurement Filter AND (Reviews OR Systematic Reviews)
limit 114 to "review articles" (549)
115 114 and 151 (195)
116 115 or 152 [All Measures limited to Reviews] (597)

(11 and 85) not 101 [SF-36 & SF-12] (6200)
limit 154 to "review articles" (215)
155 154 and 151 (93)
156 155 or 156 [SF-36 & SF-12 limited to Reviews] (241)

(15 and 85) not 101 [EQ-5D] (1153)
limit 159 to "review articles" (84)
160 159 and 151 (38)
161 160 or 161 [EQ-5D limited to Reviews] (91)

(22 and 85) not 101 [PROMIS] (102)
limit 164 to "review articles" (3)
165 164 and 151 (2)
167  165 or 166 [PROMIS Limited to Reviews] (4)

168  (27 and 85) not 101 [WHOQOL] (430)
169  limit 168 to "review articles" (18)
170  168 and 151 (4)
171  169 or 170 [WHOQOL Limited to Reviews] (18)

172  (33 and 85) not 101 [HUI] (298)
173  limit 172 to "review articles" (22)
174  172 and 151 (8)
175  173 or 174 [HUI Limited to Reviews] (23)

176  (34 and 85) not 101 [Nottingham Health Profile] (367)
177  limit 176 to "review articles" (34)
178  176 and 151 (5)
179  177 or 178 [Nottingham Health Profile Limited to Reviews] (35)

180  (40 and 85) not 101 [QWB] (146)
181  limit 180 to "review articles" (22)
182  180 and 151 (4)
184  181 or 182 [QWB Limited to Reviews] (22)

185  (43 and 85) not 101 [Hows Your Health] (0)
186  (50 and 85) not 101 [AQoL] (421)
187  limit 186 to "review articles" (76)
188  186 and 151 (15)
189  187 or 188 [AQoL Limited to Reviews] (79)

**General MeSH Subject Headings for Quality of life AND Measurement Filter AND (Reviews OR Systematic Reviews)**

190  *Quality of Health Care/ (25717)
191  (190 and 85) not 101 (2121)
192  limit 191 to "review articles" (247)
193  190 and 151 (282)
194  192 or 193 [Quality of life & measures & reviews] (466)

195  Health Status/ or Health Status Indicators/ (69412)
(195 and 85) not 101 (14935)
196
197
limit 196 to "review articles" (1301)
198
196 and 151 (350)
199
197 or 198 (1396) [Health Status/indicators & measures & reviews]
200
193 or 199 [General QoL terms & measures & reviews] (1678)
201
200 not 153 (1584)
202
limit 201 to English language (1465)
203
103 not 163 [eq-5D] (1197)
204
154 not 157 [SF - reviews] (6466)
205
159 not 163 [EQ-5D - reviews] (1197)
206
164 not 167 [PROMIS - reviews] (109)
207
168 not 171 [WHOQOL] (475)
208
172 not 175 [HUI - reviews] (301)
209
176 not 179 [Nottingham - reviews] (332)
210
180 not 184 [QWB - reviews] (126)
211
186 not 189 [AQoL - reviews] (368)
212
limit 204 to English language (6062)
Appendix 5. Search Results (for reviews of psychometric properties)

![Flowchart with search results]

**Figure 3.** Results of search for reviews of psychometric properties
### Appendix 6. Descriptive Data for the Nine Shortlisted PROMs Instruments

**Table 16. Instrument details**

<table>
<thead>
<tr>
<th>Instrument Abbreviation</th>
<th>Name</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>Assessment of Quality of Life</td>
<td>The AQoL is a multi-attribute utility measure for use in economic evaluation, measuring health-related quality of life. The descriptive system can be used to provide health profiles.</td>
</tr>
<tr>
<td>NHP</td>
<td>Nottingham Health Profile</td>
<td>To provide a brief indication of a patient's perceived emotional social and physical health problems.</td>
</tr>
<tr>
<td>SF-12® / SF-12v2™</td>
<td>SF-12® Health Survey and SF-12v2™ Health Survey</td>
<td>Developed to be a much shorter, yet valid, alternative to the SF-36 for use in large surveys of general and specific populations as well as large longitudinal studies of health outcomes. The SF-36 was developed during the Medical Outcomes Study (MOS) to measure generic health concepts relevant across age, disease, and treatment groups. The SF-12 is a subset of the SF-36.</td>
</tr>
<tr>
<td>HUI®: HUI3</td>
<td>Health Utilities Index</td>
<td>To describe health status, measure within-attribute morbidity and health-related quality of life, and produce utility scores. There are three versions of the HUI; HUI, HUI2 and HUI3.</td>
</tr>
<tr>
<td>WHOQoL-BREF</td>
<td>World Health Organization Quality of Life assessment instrument</td>
<td>To assess individuals' perceptions on the quality of their life.</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>Quality of Well-Being scale Self-Administered</td>
<td>To measure health-related quality of life, to monitor the health of populations over time, or to evaluate the efficacy and effectiveness of clinical therapies of practices using a preference-weighted self-administered measure.</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Euroqol EQ-5D</td>
<td>To assess health outcome from a wide variety of interventions on a common scale, for purposes of evaluation, allocation and monitoring. Of note, the EQ-5D-3L consists of the original EQ-5D descriptive system and the EQ visual analogue scale (EQ-VAS).</td>
</tr>
<tr>
<td>PROMIS: Global Health</td>
<td>Patient Reported Outcomes Measurement Information System</td>
<td>A system of highly reliable, valid, flexible, precise, and responsive assessment tools that measure patient-reported health status. The Global Health index refers to a group of 10 validated global health items that can be used to measure general health. These items can be used together as a global health instrument.</td>
</tr>
</tbody>
</table>
Instrument Cost Notes:


**NOTTINGHAM HEALTH PROFILE**: “No. Fees apply for commercial / pharmaceutical companies” Information retrieved Jan 3, 2013 at: [http://www.outcomesdatabase.org/node/669](http://www.outcomesdatabase.org/node/669)


**Health Utilities Index**: Health Utilities Inc. does not grant permission for copies of its proprietary materials (e.g., questionnaires) to be distributed in grant proposals or reports. The HUI Application and Interpretation Package costs $CAN 4,000.00. This package includes initial consultation to determine the most appropriate questionnaires for use in a specific study, and permission to use one version of an HUI questionnaire and the matching procedures manual set in one study. Additional questionnaires and manuals are available for use at a cost of $CAN 2,000.00 each per study. If the study requires more than one questionnaire the fee schedule becomes more complicated. For example, a study using two self-complete questionnaires (e.g., self-complete and self-assessed in both English and Spanish) should budget $CAN 6,000.00 (the manual is in common) while a study using a self-complete and an interviewer-administered questionnaire should budget $CAN 8,000.00 (fee reflects payment for one additional questionnaire and one additional manual). HUI grants permission for use of its proprietary materials (e.g., instrumentation) one study at a time. (Additional information regarding the fee
schedule is presented on the web site: http://healthutilities.biz/fees.htm . Source: Email correspondence with Mr. Bill Furlong, General Manager Health Utilities Inc, 24/09/2012.


Quality of Well-Being Scale – Self Administered (QWB-SA): Copyright Agreement For Non-Profits appears to indicate free use with restrictions. Information retrieved Jan 6, 2013 at: https://hoap.ucsd.edu/qwb-info/NotforProfit-Copyright.pdf . Terms and cost for use by non-profits indicates, “Scoring - <HSRC Scoring - $57/hr or, an algorithm for scoring can be purchased for $240. The scoring instructions are free of charge. Information retrieved Jan 6, 2013 at: https://hoap.ucsd.edu/qwb-info/price-nonprofit.aspx


EQ-5D: “Licensing fees are determined by the EuroQol Executive Office on the basis of the user information provided on the registration form. The amount is dependent upon the type of study/trial/project, funding source, sample size and number of requested languages. You are not obligated to purchase by registering.” Information retrieved Jan 3, 2013 at: http://www.euroqol.org/eq-5d-products/how-to-obtain-eq-5d.html The User Guide can be downloaded from www.euroqol.org homepage.

PROMIS Short Form v1.0 - Global Health Scale can be obtained for free by email request. Information retrieved Jan 3, 2013 at: https://www.assessmentcenter.net/PromisForms.aspx . While it appears use of the GHS is free, users must “agree that the PROMIS Health Organization and PROMIS Cooperative Group provide access to PROMIS instruments (e.g., item banks, short forms, profile measures) subject to the PROMIS Terms and Conditions (PTAC).” Information retrieved Jan 3, 2013 at: http://www.assessmentcenter.net/documents/PROMIS%20Terms%20and%20Conditions%20v7_3.pdf
### Table 17. Age groups instruments have been adapted for

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Population of Intended Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQoL</td>
<td>Adult (14+ years)</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Adult [ +12 years, child-friendly versions in development ]</td>
</tr>
<tr>
<td>HUI</td>
<td>All (5+ years)</td>
</tr>
<tr>
<td>NHP</td>
<td>Adult (16 + years)</td>
</tr>
<tr>
<td>SF-36</td>
<td>Adult; Adolescent (14+ years)</td>
</tr>
<tr>
<td>SF-12</td>
<td>Adult; Adolescent (14+ years)</td>
</tr>
<tr>
<td>QWB-SA</td>
<td>Adult (18+)</td>
</tr>
<tr>
<td>WHOQoL</td>
<td>Adult (18+)</td>
</tr>
<tr>
<td>PROMIS/GHS</td>
<td>Differential item functioning (DIF) in key groups (age, sex, diagnostic grouping, education, race/ethnicity, language translation)</td>
</tr>
</tbody>
</table>
Appendix 7. Summary of the Psychometric Evidence

Table 18. Legend for the summary of the psychometric evidence

<table>
<thead>
<tr>
<th>Strength of Evidence</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>+++ or ---</td>
</tr>
<tr>
<td>Moderate</td>
<td>++ or --</td>
</tr>
<tr>
<td>Limited</td>
<td>+ or -</td>
</tr>
<tr>
<td>Conflicting</td>
<td>+/-</td>
</tr>
<tr>
<td>Reported evidence does not pertain to COSMIN criteria</td>
<td>any combination of symbols, followed by /?</td>
</tr>
<tr>
<td>Unknown</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 19. Summary of the psychometric evidence of the “A-List” (Top three instruments selected by January 14, 2013 workshop participants)

<table>
<thead>
<tr>
<th>Psychometric Property</th>
<th>EQ-5D</th>
<th>SF-36</th>
<th>PROMIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal Consistency</td>
<td>N/A, Not a reflective measure.</td>
<td>++++, All SF-36 scales demonstrated acceptable internal consistency, with Cronbach's alpha ranging from 0.8 - &gt;0.92 and evidence of unidimensionality. However, there is some mention of some subscales mapping onto both of the 2 primary components (mental and physical health), suggesting that they are not entirely orthogonal.</td>
<td>++++, Internal consistency greater than .8 for both subscales and the whole instrument. Unidimensionality of the subscales was in part confirmed by the factor analysis..</td>
</tr>
<tr>
<td>Reliability</td>
<td>+/-, Mixed evidence from test-retest. Some studies suggest subcomponents are below the COSMIN standards. For the 5 dimensions, the k coefficients ranged from 0.63 to 0.80. The reliability coefficient in the rheumatoid arthritis patients was 0.55.</td>
<td>++++, For the two composite scores, +/- for the individual subscales. Most correlations were above 0.7 and 0.8 but some fell below to 0.63.</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td>Psychometric Property</td>
<td>EQ-5D</td>
<td>SF-36</td>
<td>PROMIS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Content Validity</strong></td>
<td>- Several studies have found that its coverage is less than complete, which is the trade-off for its short length. The EQ-5D was found to cover 35.9% of the health items, prone to large percentages at ceiling with less than severe illness.</td>
<td>++, Good evidence that it is comprehensive. Overall, it is perceived to be a good generic instrument, which is relevant to general population.</td>
<td>+/-?, Comprehensive nature was part of the item selection, but no direct tests with population of how valid it was for them.</td>
</tr>
<tr>
<td><strong>Construct Validity</strong></td>
<td>+/-?, Not specifically described</td>
<td>++++, Overall, strong support for 2 factor solution, but one study did report conflicting results of a confirmatory factor analysis. Good evidence for the factor structure being confirmed.</td>
<td>+/-/? Nothing reported directly on the proportion of variance explained by the 2 proposed factors, but the two composite scores do explain 60% of the variance in EQ-5D scores. Also found to be correlated with other item scales from PROMIS.</td>
</tr>
<tr>
<td><strong>Cross-cultural Validity</strong></td>
<td>+/-?, There is some evidence to support it as being valid in a variety of contexts, given the way it was developed and translated, but nothing that speaks directly to COSMIN criteria.</td>
<td>+/-/-?, No report of actual factor structure for evaluating based on COSMIN criteria. Some indication that there may be issues, but the reviews were short on specifics. Certain groups loaded higher on mental component than others.</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td><strong>Criterion Validity</strong></td>
<td>+/-, Absence of gold standard is problematic, but correlations with SF-36 and SF-12 did not meet COSMIN criteria for all subscales, Moderate correlation with SF-36 scales(r=.36 to .80, except emotion role scale (r=.09)</td>
<td>?, not specifically described. Note: No gold standard for comparison is a problem, and not much detail on correlation with other instruments reported in the review.</td>
<td>? Absence of gold standard, but was significantly correlated with EQ-5D scores. R-squared GPH=.82, GMH=.61</td>
</tr>
<tr>
<td>Psychometric Property</td>
<td>EQ-5D</td>
<td>SF-36</td>
<td>PROMIS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>--/?</td>
<td>+++/?, Strong evidence in other forms, that it is highly responsive, correlated with changes in clinical outcomes, and correlated with known groups, but nothing that speaks directly to COSMIN criteria.</td>
<td>++, Good evidence based on correlations with EQ-5D scores, which was stronger for physical health than for mental health, and correlation with pain impact and depressive symptoms.</td>
</tr>
<tr>
<td><strong>Generalizability and Interpretability</strong></td>
<td>Seems good for discriminating between known groups, but problems with using it in populations with good health.</td>
<td>Generally demonstrated, strong evidence of correlation with clinical change scores, good evidence of responsiveness.</td>
<td>Not specifically described</td>
</tr>
<tr>
<td><strong>Comparison with Other Instruments</strong></td>
<td>Short and quick, but less sensitive than other instruments, and not as comprehensive as others.</td>
<td>Consistently reported to be better than other instruments in terms of reliability, responsiveness, and validity. More responsive than NHP.</td>
<td>Similar internal consistency to SF-36, but 2 minutes to complete versus 10 minutes for SF-36.</td>
</tr>
</tbody>
</table>

Table 20. Summary of the psychometric evidence of the “B-List” (Three instruments eliminated from the final list in the second cut by January 14, 2013 workshop participants)

<table>
<thead>
<tr>
<th>Psychometric Property</th>
<th>AQoL</th>
<th>HUI</th>
<th>WHOQoL-BREF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal Consistency</strong></td>
<td>+/-?, Although it is reported to have good unidimensionality no reporting of alpha estimates or degree to which subscales are unidimensional.</td>
<td>?, not specifically described</td>
<td>+/-?, Values for Cronbach's alpha were acceptable (&gt;0.7) for physical health (0.82), psychological (0.81), environment 0.80, but marginal for social relationships (0.68).</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>?, Not specifically described</td>
<td>+/-/-?, 8 of the 10 individual questions and 6 of the 8 attributes had moderate or better kappa coefficients. Note: reported as moderate and not the exact Kappas.</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td>Psychometric Property</td>
<td>AQoL</td>
<td>HUI</td>
<td>WHOQoL-BREF</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------</td>
<td>-----</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>Content Validity</strong></td>
<td>+/-, Information indicating AQoL provides broad coverage, but authors indicate that they did a literature search for articles which dealt with content validity and didn't turn up any articles for AQoL.</td>
<td>+/-/-?, Seems to be comprehensive for some aspects of health, but lacking on mental health. Strong support for predictive validity of scores.</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td><strong>Construct Validity</strong></td>
<td>?, Not specifically described</td>
<td>?, Not specifically described</td>
<td>++++, Factor structure confirmed, and also stated to be fairly stable across sites.</td>
</tr>
<tr>
<td><strong>Cross-cultural Validity</strong></td>
<td>?, Not specifically described</td>
<td>+/-?, Widespread international use, and language versions suggest strong cross cultural validity, but nothing reported speaking to COSMIN criteria.</td>
<td>+++ Very detailed evidence on instrument construction, and cross country validity.</td>
</tr>
<tr>
<td><strong>Criterion Validity</strong></td>
<td>?, Not specifically described</td>
<td>?, Absence of a gold standard instrument. No correlations reported in reviews.</td>
<td>? Absence of a gold standard instrument. No correlations reported in reviews.</td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>+/-?, Progressive deterioration of scores between different groups of patient populations (i.e. community, out-patient, vs. in-patient) was noted, but nothing directly related to COSMIN criteria.</td>
<td>+/-/-?, Noting pertaining directly to COSMIN criteria. Citations of studies which demonstrate its ability to distinguish between known groups, but also has ceiling effect issues (&gt;20% had perfect score in general population sample).</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td><strong>Generalizability and Interpretability</strong></td>
<td>Scores are correlated with known groups of patients (e.g. in-patient vs. outpatient)</td>
<td>Widely used and scores correlated in expected direction, with known groups. However, a study is cited in which scores were not correlated between patients with acute problem and a general population sample. Only 2 of 21</td>
<td>Good evidence that scores are correlated with known ill and well patient populations.</td>
</tr>
</tbody>
</table>
studies considered in one of our review papers, which used this instrument, were conducted outside McMaster.

<table>
<thead>
<tr>
<th>Comparison with Other Instruments</th>
<th>AQoL</th>
<th>HUI</th>
<th>WHOQoL-BREF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not specifically described</td>
<td>HUI has a ceiling effect, and is not recommended if trying to distinguish among those in good health.</td>
<td>Not specifically described</td>
<td></td>
</tr>
</tbody>
</table>

Table 21. Summary of the psychometric evidence of the “C-List” (First two instruments to be eliminated from the final list in the first cut by January 14, 2013 workshop participants)

<table>
<thead>
<tr>
<th>Psychometric Property</th>
<th>NHP</th>
<th>QWB-SA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal Consistency</strong></td>
<td>+/-, Overall internal consistency meets criteria, but not for all subscales. Also, nothing reported on unidimensionality.</td>
<td>N/A, Not a reflective measure.</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>+/-, Test-retest ranged from .7 to .92, which is both above and below the COSMIN standard.</td>
<td>?, While one study states that it has 1 day test-retest reliability, data was gathered during the same interview, asking about health the day prior and today. This is a major methodological problem for assessing reliability.</td>
</tr>
<tr>
<td><strong>Content Validity</strong></td>
<td>+/-?, The way it was developed suggests it is comprehensive, but not clear that it is applicable outside of populations with major health issues.</td>
<td>+/-, Physical health seems to be comprehensive, but lacking on mental health.</td>
</tr>
<tr>
<td><strong>Construct Validity</strong></td>
<td>+/-, Correlation coefficient between dimension scores and the Global Health Assessment ranged from 0.29 to 0.49.</td>
<td>?, Not specifically described</td>
</tr>
<tr>
<td><strong>Cross-cultural Validity</strong></td>
<td>+/-?, Nothing reported on factor structure. Description of development indicates potential for cross-cultural equivalence, but some issues identified with item weights in different cultural contexts.</td>
<td>?, Multiple language versions exist, but validity has not been examined in these different contexts.</td>
</tr>
<tr>
<td><strong>Criterion Validity</strong></td>
<td>? Absence of a gold standard instrument. No correlations reported in reviews.</td>
<td>? No gold standard for criterion validity, but it is reported to be ‘significantly’ correlated with SIP and SF-36.</td>
</tr>
<tr>
<td>Psychometric Property</td>
<td>NHP</td>
<td>QWB-SA</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>?/-, Not very responsive for healthy individuals. Issues with ceiling effects as well, but nothing reported which speaks to the COSMIN scoring criteria.</td>
<td>+/-, Met the COSMIN criteria for general health, but not all aspects. Stated problems in dealing with mental health, and problems with over weighting minor conditions (like wearing eyeglasses).</td>
</tr>
<tr>
<td><strong>Generalizability and Interpretability</strong></td>
<td>Large evidence base that shows it is not very useful in healthy populations. Widespread use in the UK, and seems to have a niche of being used with moderately to severely ill populations.</td>
<td>Strong evidence reported for its generalizability in some papers, but other raised issues with the structure of the instrument and the dichotomous response format for generating meaningful data that captures change. Differential coverage of certain aspects of health is also noted, such that wearing glasses can more negatively impact score than other health conditions.</td>
</tr>
<tr>
<td><strong>Comparison with Other Instruments</strong></td>
<td>Not as sensitive as SF-36, and not well suited for a general population, but may be better in populations with poor health. Evidence that scores are more reproducible, but less responsive than other instruments.</td>
<td>Good for avoiding ceiling effects, but doesn't outperform other instruments in any other ways. Response burden is an issue as well.</td>
</tr>
</tbody>
</table>
Appendix 8. Quality assessment of the articles reviewed (psychometric review)

Table 22. Quality assessment of the articles reviewed

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Reviews (Total=22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provided a rational for review</td>
<td>22</td>
</tr>
<tr>
<td>Provided an explicit statement of what questions were being addressed</td>
<td>21</td>
</tr>
<tr>
<td>Provided an interpretation of the evidence in the context of what is known</td>
<td>21</td>
</tr>
<tr>
<td>Summarized main findings in relation to questions being addressed</td>
<td>20</td>
</tr>
<tr>
<td>Review article comparing multiple instruments</td>
<td>10</td>
</tr>
<tr>
<td>Discussed limitations of the review</td>
<td>6</td>
</tr>
<tr>
<td>Described funding source</td>
<td>5</td>
</tr>
<tr>
<td>Described literature search strategy</td>
<td>3</td>
</tr>
<tr>
<td>Described inclusion criteria</td>
<td>3</td>
</tr>
<tr>
<td>Provided full electronic search strategy</td>
<td>2</td>
</tr>
<tr>
<td>Provided a statement about risk of bias across studies</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix 9. January 14, 2013 PROMS Workshop Participants

The following stakeholders participated, with the research team, in the Patient Reported Outcome Measurement (PROMS) and Integrated Primary and Community Care (IPCC) Workshop in Victoria, BC on January 14, 2013.

**CIHR**

Erik Landriault  
Assistant Director  
CIHR Institute of Health Services and Policy Research

**BC Ministry of Health**

Barbara Korabek  
Assistant Deputy Minister  
Health Authorities Division

Victoria Schuckel  
A/Executive Director  
Lead – MoH-wide MELS planning

Natacha Vairo  
Policy Analyst  
MoH-wide MELS planning

Barbara Smith  
Senior Policy Analyst  
Research Branch

Kelly Barnard  
Senior Medical Consultant  
Expert Advisory Group

David Gray  
Director  
Modeling and Analysis Branch

Linda Low  
Project Manager  
IPCC Medical Services and HHR -PHC

Betty Weber  
Project Director  
IPCC HAD – Home Community and Integrated Care

Ramani Kumar  
Director  
IPCC Home Community and Integrated Care

Ross Hayward  
Director  
Mental Health and Substance Use

Caryl Harper  
Manager  
Patients as Partners

Nancy South  
Director  
Performance Measurement, Analysis and reporting

Shirley Wong  
Director  
Health Sector IM/IT

Sherry Bar  
Research Officer  
Primary Health Care

Jessica Finnerty  
Director  
Patient Safety and Care Quality

James Watson  
Director  
Patient Safety and Care Quality

Lettita Pengeley

**BC Health Authority Evaluators**

Sherri Tilotson  
Northern Health  
MELS
Oluseyi Oyedele  Interior Health  MELS
George Tien  Fraser Health  MELS
Laurie Ringaert  Vancouver Coastal Health  MELS
Shelley Tice  Vancouver Island Health  MELS

BC Health Authorities

Linda Foley  Fraser  Implementation Leadership Committee
Phil Lawrence  Vancouver Island  Implementation Leadership Committee
Jana Archer  Vancouver Island
Donna Conway  Vancouver Island  Director, Performance Monitoring & Reporting
Katy Mukai
Carol Park  Vancouver Coastal  Implementation Leadership Committee
Mark Chase  Vancouver Coastal  Executive Director, Decision Support
Diane Kostachuck  Interior  Implementation Leadership Committee
Glenn Kissmann  Interior  Director, Strategic information
G Miller  PHA

Michael Smith Foundation for Health Research

Trilby Smith  Evaluation Lead (MELS)  MELS

Others

Lena Cuthbertson  Providence Health Care
Liza Kallstrom  Practice Supports Program
Christina Southey  Impact BC
Dr. Patrick McGowan  Associate Professor  UVIC
O Djurgev  PHSA
Juanita Arthur
Appendix 10. January 14, 2013 Workshop Evaluation Results

Workshop: Patient reported outcome measurement (PROMs) and integrated primary and community care (IPCC)

Victoria, BC  January 14, 2013

Evaluation Summary:

- Response rate: 28/40
- The participants most strongly agreed that they better understood the basics of PROMS, and that the presenters were knowledgeable.
- While they agreed that the workshop helped them understand the IPCC policy context for the use of PROMs in BC; and PROM instruments and their potential use in IPCC projects – these two questions received the lowest ratings and 5 of the 28 respondents disagreed that the workshop helped them understood the IPCC context for the use of PROMS in BC
- 17 participants commented on what they most enjoyed about the workshop, with the overall theme being that they liked the:
  - networking, interaction, information, group discussions
- 14 participants made suggestions to improve the workshop, with the overall theme being that they would have liked:
  - information ahead of time; larger print handouts; more discussion and information on instruments, context, objectives, long term plans, and implementation strategies.
- 13 participants wrote additional comments and specific requests for the research team included:
  - report back to participants re: next steps and further progress made
  - Please email out power point used.
  - Could we create a tool kit of info you have collected re: 9 tools…it feels that we should capitalize on info learned on all these tools.
  - Needed more info re PROMIS

Participants were asked: “We would like to know the extent to which this workshop has met your expectations. Please indicate your agreement or disagreement with each statement and provide further comments if you wish. (Please circle your selected response.)"
Question scale: 4 - strongly agree; 3 - agree; 2 - disagree; 1 - strongly disagree

<table>
<thead>
<tr>
<th>Questions</th>
<th>average score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The content of the workshop was useful.</td>
<td>3.5</td>
</tr>
<tr>
<td>2. Given the time allowed, the amount of material covered was appropriate.</td>
<td>3.3</td>
</tr>
<tr>
<td>3. The learning objectives were met.</td>
<td>3.1</td>
</tr>
<tr>
<td>4. What I have learned in this workshop will help me understand:</td>
<td></td>
</tr>
<tr>
<td>- the basics of PROMs</td>
<td>3.8</td>
</tr>
<tr>
<td>- the IPCC policy context for the use of PROMs in BC</td>
<td>3.0</td>
</tr>
<tr>
<td>- PROM instruments and their potential use in IPCC projects</td>
<td>3.2</td>
</tr>
<tr>
<td>5. The presenters were knowledgeable about the subject.</td>
<td>3.9</td>
</tr>
<tr>
<td>6. The materials were presented in an organized way.</td>
<td>3.5</td>
</tr>
<tr>
<td>7. Appropriate audiovisuals were used during the presentation.</td>
<td>3.4</td>
</tr>
<tr>
<td>8. Overall, I am satisfied with this workshop.</td>
<td>3.4</td>
</tr>
</tbody>
</table>
References


Group; 2011 Apr [cited 2012 Sep]. Available from:

53. Hawthorne G, Richardson J. Measuring the value of program outcomes: A review of
multiattribute utility measures. Expert Review of Pharmacoeconomics and Outcomes

Available from:

55. PROMIS. Instrument development and psychometric evaluation: scientific standards
[internet]. National Institute of Health; 2012 May [cited 2013 Apr 1]. Available from:
http://www.nihpromis.org/science/publications

56. PROMIS. Calibration Testing [internet]. [cited 2012 Nov]. Available from:
http://www.nihpromis.org/science/calibrationtesting

57. PROMIS. Instrument Details [internet]. [cited 2012 Nov]. Available from:
http://www.nihpromis.org/measures/instrumentdetails

http://howsyourhealth.org/.

32(4) [cited 2012 Sep]. Available from:
http://dartmed.dartmouth.edu/summer08/html/vs_website.php

60. Galen Research. Nottingham Health Profile: Guidelines for users [online] 2012 Apr. Email to
Janice Murphy (immurphy@shaw.ca) 2012 Sept 29 [cited 2012 Dec 18].

61. Health Utilities Index (HUI). Email to Janice Murphy (immurphy@shaw.ca) 2012 Sept 24
[cited 2012 Dec 18].

PubMed search filter for finding studies on measurement properties of measurement

63. Scottish Intercollegiate Guidelines Network. SEARCH FILTERS [Internet].[Updates 2012

statement for reporting systematic reviews and meta-analyses of studies that evaluate


