Performing a systematic review of measurement properties

Riekie de Vet¹
Caroline Terwee¹
Wieneke Mokkink¹
Donald Patrick²

¹ VU University Medical Center
Knowledgecenter Measurement Instruments
Department of Epidemiology and Biostatistics

² Department of Health Services
Seattle Quality of Life Group\Center for Disability Policy and Research at the University of Washington
Content of the workshop

1. Presentation:
   - Introduction: aim and importance of systematic reviews of measurement properties
   - 10-step procedure for performing a systematic review of measurement properties
   - Example: systematic review of the measurement properties of questionnaires measuring pain and disability in patients with neck pain
   - Discussion points

2. Exercise: data synthesis: combining evidence on measurement properties from different studies
What is a systematic review of measurement properties?

A systematic review of all available studies on the measurement properties of one or more measurement instruments, that aim to measure a specific construct of interest...

...in which the quality of the studies and the measurement properties are being evaluated and compared according to predefined quality criteria.
Why a systematic review of measurement properties?

- Often there are many instruments for measuring a specific construct.
- We want to choose the best available instrument for a specific purpose, for research or clinical practice.
- Measurement properties are an important factor in choosing an instrument. Using instruments with good measurement properties will lead to detection of smaller treatment effects, more power to draw strong conclusions, better interpretation of study results.
Why a systematic review of measurement properties?

- Researchers and clinicians do not have time and expertise to find all instruments and to select the best one for a specific purpose.

- High-quality reviews provide evidence for the best instrument for a specific purpose.
“This review provides information for researchers and clinicians to facilitate the choice among the existing questionnaires for shoulder disability.

The DASH and SPADI are recommended for evaluative purposes in outpatient clinics. The SIQ is recommended for evaluation of patients with shoulder instability and the OSQ for evaluation of patients having a shoulder operation other than stabilisation. For discriminative purposes the SST is suggested for patients with shoulder complaints in general. The SSRS and SPADI should not be used for assessment of individual patients.”

Why a systematic review of measurement properties?

- A systematic review of measurement properties provides a quality rating of each available instrument.

Quality ratings might be useful for e.g. Cochrane reviewers to evaluate the quality of trials (outcome measures).

The Cochrane PRO methods group distributes reviews of measurement properties among Cochrane review groups.

Quality ratings might be useful for editors, grant suppliers, FDA, EMEA, etc.
Why a systematic review of measurement properties?

• To set a research agenda for further validation of promising instruments

• To decide on the necessity to develop a new instrument
“In this review 23 of the 85 included instruments were rated as having sufficient content validity. We recommend validating these questionnaires further for reliability, construct validity, and especially responsiveness.”

In recent studies, the IPAQ seems to be used most often and it is by far the most widely validated questionnaire at present. Reliability of the IPAQ was not shown consistently within or between studies. Validity of the IPAQ seems questionable. Responsiveness has not been assessed. Therefore, we feel that additional well designed studies on the measurement properties, with specific attention to responsiveness, of the IPAQ are required.”

Published systematic reviews
Quality of systematic reviews


Review of 148 reviews

• 22% only searched one database
• search strategy often poorly described and often incomplete
• often not reported whether article selection was performed by two reviewers
• methodological quality of the included studies mostly not assessed
• data synthesis is different in each review
• no standardized methodology
Systematic reviews of measurement properties are quite complex

- The review has more than one outcome measure, i.e. multiple measurement properties

- Different studies evaluate different measurement properties → the number of studies in the analysis is different per measurement property

- The quality of the studies is evaluated per measurement property

- Data synthesis is different per measurement property

- Evidence for one measurement property may come from different studies

Therefore, a systematic review of measurement properties is actually a collection of separate reviews per measurement property
Performing a systematic review of measurement properties

10 step procedure to ...

- Identify all available instruments
- Summarize the evidence on the measurement properties

Based on methodology of the Cochrane Collaboration
10 steps in a systematic review of meas prop

1. Formulating a research question
2. Performing a literature search
3. Formulating eligibility criteria
4. Selecting abstracts and full-text articles
5. Evaluating the methodological quality of the included studies
6. Extracting the data
7. Comparing the content
8. Data synthesis- evaluating the evidence for adequacy of the measurement properties
9. Drawing an overall conclusion of the systematic review
10. Reporting on the systematic review
1. Formulating the research question

Different kind of reviews

- a review of all available studies on the measurement properties of one measurement instrument

- a review of all available studies on the measurement properties of a selection of the most commonly used measurement instruments that aim to measure a particular construct in a particular population

- a review of all available studies on the measurement properties of all available measurement instruments that aim to measure a particular construct in a particular population

- a review of all available studies on the measurement properties of all available measurement instruments (without specifying the construct of interest) in a particular population
1. Formulating the research question

Different kind of reviews

• a review of all available studies on the measurement properties of one measurement instrument

• a review of all available studies on the measurement properties of a selection of the most commonly used measurement instruments that aim to measure a particular construct in a particular population

• a review of all available studies on the measurement properties of all available measurement instruments that aim to measure a particular construct in a particular population

• a review of all available studies on the measurement properties of all available measurement instruments (without specifying the construct of interest) in a particular population
1. Formulating the research question

The research question includes four key elements

1. construct of interest
2. target population
3. type of measurement instrument
4. measurement properties
Objective:

“to critically appraise and compare the measurement properties of neck-specific questionnaires measuring pain and disability in patients with non-specific neck pain”
Objective:

“to critically appraise and compare the measurement properties of neck-specific questionnaires measuring pain and disability in patients with non-specific neck pain”
2. Performing a literature search

- Use at least two databases: Medline, Embase, Cinahl, PsycINFO ...

- Formulate search terms for
  1. construct of interest
  2. target population
  3. type of measurement instrument
  4. measurement properties
Databases:

- PubMed (1966 to October 2009)
- EMbase (1974 to October 2009)
- Cinahl (1981 to October 2009)
- PsycINFO (1806 to October 2009)
Example PubMed search terms

1. construct of interest: pain and disability/functioning
No specific search terms (see target population)

2. target population: non-specific neck pain

("Neck"[Mesh] OR "Neck Pain"[Mesh] OR "Neck Injuries"[Mesh])

3. Type of measurement instrument: questionnaires

For finding PROs

Search filter, developed by the Patient Reported Outcomes Measurement Group (Oxford)

http://phi.uhce.ox.ac.uk/
Elizabeth.Gibbons@dphpc.ox.ac.uk

Was not available yet for this review
4. Measurement properties

PubMed search filter for finding studies on measurement properties:
Terwee et al. QLR 2009;18:1115-1123


NOT (("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type] OR "animals"[MeSH Terms] NOT "humans"[MeSH Terms])
3. Formulating eligibility criteria

Inclusion: the article concerns:

1. instruments measuring the construct of interest
2. a study sample representing the target population of interest
3. the type of measurement instrument of interest
4. development of a measurement instrument or the evaluation of one or more of the measurement properties of an instrument
1. Instrument aims to measure pain and/or disability \ / physical functioning

2. Neck pain is the main complaint of the study population

3. Instrument is a self-report questionnaire specifically developed or adapted for patients with neck pain

4. Aim of the study is the development or evaluation of the measurement properties of a neck-specific questionnaire

5. Full-text article, published in English
Points of discussion

1. Should we include studies of all languages?

2. Should we include other kind of studies, e.g. studies in which an instrument of interest is used as an outcomes measure (e.g. clinical trials or other longitudinal studies), where the results of the trial may provide evidence for the responsiveness of the instrument?
4. Selecting relevant articles

- Using the inclusion criteria, based on title and abstract or full text article
- Consider eligibility of the article
- Selection performed by two researchers, independently
- Document reasons for exclusion
- Present a flow chart, describing the search and selection process
- Check references for additional studies
3424 records were found with the search
110 eligible full-text articles were screened
46 articles were included
8 questionnaires were identified:

- Neck Disability Index (NDI) (23 studies)
- Neck Pain and Disability Scale (NPDS) (13 studies)
- Neck Bournemouth Questionnaire (NBQ) (4 studies)
- Northwick Park Neck Pain Questionnaire (NPQ) (10 studies)
- Copenhagen Neck Functional Disability Scale (CNFDS) (4 studies)
- Core Neck Questionnaire (CNQ) (2 studies)
- Whiplash Disability Questionnaire (WDQ) (3 studies)
- Core Whiplash Outcome Measure (CWOM) (1 study)
Articles retrieved by search strategy (n = 3424)

Articles selected based on title and abstract (n = 110)

Articles selected based on full text (n = 46)

Articles included in review (n = 46)†:
•NDI (n=23)
•NPDS (n=13)
•NBQ (n=4)
•NPQ (n=10)
•CNFDS (n=4)
•CNQ (n=2)
•WDQ (n=3)
•CWOM (n=1)

Main reason for exclusion:
•article not retrievable (n=2)
•not full text original article (n=7)
•validation not aim of study (n=19)
•neck pain not main complaint (n=14)
•specific neck disorder (n=4)
•not neck-specific questionnaire (n=18)

† The sum of the different questionnaires is higher than 46, because some studies evaluate more than one questionnaire
5. Evaluating the methodological quality of the included studies

COSMIN Checklist

Checklist to rate the methodological quality of an article evaluating measurement properties

COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments
5. Evaluating the methodological quality of the included studies

The COSMIN checklist consists of 9 different boxes, one for each measurement property, with questions regarding quality aspects of the design and statistical analysis.

www.cosmin.nl
1. Was the percentage of missing items given?
2. Was there a description of how missing items were handled?
3. Was the sample size included in the analysis adequate?
4. Were at least two measurements available?
5. Were the administrations independent?
6. Was the time interval stated?
7. Were patients stable in the interim period on the construct to be measured?
8. Was the time interval appropriate?
9. Were there any important flaws in the design or methods of the study?

10. for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?
11. for dichotomous/nominal/ordinal scores: Was kappa calculated?
12. for ordinal scores: Was a weighted kappa calculated?
13. for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic
14. Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions
• For reviews on measurement properties a 4-point rating scale was developed

• Each item is scored as excellent, good, fair, or poor

• A quality score per box is calculated by taking the lowest rating of any item in the box ("worse score counts")
Example

COSMIN scores for 46 studies on measurement properties of neck-specific questionnaires
6. Extracting the data

Which data should be extracted?

1. Characteristics of the instrument, such as construct, number of items, response options, interpretation

2. Characteristics of the population, such as age, gender, important disease characteristics

3. Results of the measurement properties
### Characteristics of the questionnaires

<table>
<thead>
<tr>
<th></th>
<th>NDI</th>
<th>NPDS</th>
<th>NBQ</th>
<th>NPQ</th>
<th>CNFDS</th>
<th>CNQ</th>
<th>WDQ</th>
<th>CWOM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Construct</strong></td>
<td>Pain, physical functioning</td>
<td>Pain, physical functioning, psychological functioning</td>
<td>Pain, physical functioning, social functioning, psychological functioning</td>
<td>Symptoms, physical functioning, social functioning, psychological functioning</td>
<td>symptoms, physical functioning, social functioning, psychological functioning</td>
<td>symptoms, physical functioning, satisfaction</td>
<td>Pain, physical functioning, psychological functioning</td>
<td>Pain, physical functioning, psychological functioning</td>
</tr>
<tr>
<td><strong>Target population</strong></td>
<td>Neck pain</td>
<td>Neck pain</td>
<td>Non-specific neck pain</td>
<td>Mechanical neck pain</td>
<td>Neck pain</td>
<td>Mechanical neck pain</td>
<td>whiplash associated disorder</td>
<td>whiplash associated disorder</td>
</tr>
<tr>
<td># questions</td>
<td>10</td>
<td>20</td>
<td>7</td>
<td>9</td>
<td>15</td>
<td>7</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td># scales</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Response options</td>
<td>1-6</td>
<td>VAS</td>
<td>0-10</td>
<td>0-4</td>
<td>0-2</td>
<td>1-5</td>
<td>0-10</td>
<td>1-5</td>
</tr>
<tr>
<td>Recall period</td>
<td>Currently</td>
<td>Currently</td>
<td>Past week</td>
<td>Currently</td>
<td>Currently</td>
<td>Past week</td>
<td>Currently</td>
<td>Past week</td>
</tr>
</tbody>
</table>
Who should decide which constructs are measured?

- Use the description of the developers?
- Define for yourself based on the content of the instrument?
### Interpretation of scores of the questionnaires

#### Minimal Important Change

<table>
<thead>
<tr>
<th>NDI</th>
<th>NPDS</th>
<th>NBQ</th>
<th>NPQ</th>
<th>CNFDS</th>
<th>CNQ</th>
<th>WDQ</th>
<th>CWOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.8 (0-50)</td>
<td>10.2 (0-50)</td>
<td>4.7 (0-50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 (0-100)</td>
<td>6-18 (0-70), 24-58%</td>
<td>6.7 (0-100)</td>
<td></td>
<td></td>
<td></td>
<td>18 (0-130)</td>
<td></td>
</tr>
</tbody>
</table>
Characteristics of the study populations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age</td>
<td>38</td>
<td>41</td>
<td>38</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>% males</td>
<td>58</td>
<td>63</td>
<td>?</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>Disease</td>
<td>Acute neck sprain</td>
<td>Cervical radiculopathy</td>
<td>Neck pain for at least 3 months</td>
<td>New complaint of non-specific neck pain</td>
<td>Chronic neck pain</td>
</tr>
<tr>
<td>Duration disease</td>
<td>1-4 days</td>
<td>25 days</td>
<td>228 days</td>
<td>45% more than 7 weeks</td>
<td>&gt;3 months</td>
</tr>
<tr>
<td>Setting</td>
<td>Physiotherapy department university hospital</td>
<td>Physiotherapy department private hospital</td>
<td>University School of Physical Therapy and Rehabilitation</td>
<td>Chiropractic practices</td>
<td>Hospital Rheumatology department</td>
</tr>
<tr>
<td>Country</td>
<td>Sweden</td>
<td>India</td>
<td>Turkey</td>
<td>UK</td>
<td>Switzerland</td>
</tr>
<tr>
<td>Response rate</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>82%</td>
</tr>
</tbody>
</table>
## Results of the measurement properties: internal consistency of the NDI

<table>
<thead>
<tr>
<th>reference</th>
<th>language</th>
<th>n</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gay et al. 2007</td>
<td>English</td>
<td>23</td>
<td>0.72-0.77</td>
</tr>
<tr>
<td>Hains et al. 1998</td>
<td>English</td>
<td>237</td>
<td>0.92</td>
</tr>
<tr>
<td>Stratford et al. 1999</td>
<td>English</td>
<td>50</td>
<td>0.87</td>
</tr>
<tr>
<td>Vernon et al. 1991</td>
<td>English</td>
<td>48</td>
<td>0.76-0.84</td>
</tr>
<tr>
<td>Nieto et al. 2008</td>
<td>Catalan</td>
<td>150</td>
<td>0.70-0.87</td>
</tr>
<tr>
<td>Trouli et al. 2008</td>
<td>Greek</td>
<td>59</td>
<td>0.85</td>
</tr>
<tr>
<td>Mousavi et al. 2007</td>
<td>Iranian</td>
<td>135</td>
<td>0.88</td>
</tr>
<tr>
<td>Lee et al. 2006</td>
<td>Korean</td>
<td>180</td>
<td>0.92</td>
</tr>
<tr>
<td>Kovacs et al. 2008</td>
<td>Spanish</td>
<td>50</td>
<td>0.89</td>
</tr>
<tr>
<td>Kose et al. 2007</td>
<td>Turkish</td>
<td>102</td>
<td>0.90</td>
</tr>
</tbody>
</table>
7. Comparing the content

- Especially important for multi-item instruments

- Which aspects are covered by the instruments (on item level)?
B Wiitavaara, M Björklund, C Brulin, M Djupsjöbacka.

How well do questionnaires on symptoms in neck-shoulder disorders capture the experiences of those who suffer from neck-shoulder disorders? A content analysis of questionnaires and interviews.

BMC Musculoskeletal Disorders 2009, 10:30
Table 4: Overview of all symptoms included in the questionnaires

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Focus</th>
<th>Neck/shoulder symptoms</th>
<th>Symptoms from the rest of the body</th>
<th>Mental/Cogn. Engagement</th>
<th>Emotional engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck Pain and Disability Scale (NPDS)</td>
<td>Neck</td>
<td>Neck pain</td>
<td>Neck stiffness</td>
<td>Sleeping difficulties</td>
<td>Affected ability to think or concentrate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Changed outlook on life (depression, hopelessness etc.) Affected emotions</td>
</tr>
<tr>
<td>Northwick Park Neck Pain Questionnaire (NPQ)</td>
<td>Neck + arms</td>
<td>Neck pain</td>
<td>Arm pain, Pins and needles in arms.</td>
<td>Sleeping difficulties</td>
<td></td>
</tr>
<tr>
<td>Patient-Specific Functional Scale Self-Reports with Neck-Dysfunction (PSFS)</td>
<td>General, tested pt with neck dysfunction.</td>
<td>Neck pain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile Fitness Mapping questionnaires (PFM)</td>
<td>Neck + arm-hand</td>
<td>Soreness, Neck pain</td>
<td>Stiffness, Tension, Cracks, Tiredness, Weakness, Lockings</td>
<td>Disturbance of balance, Dizziness, Indisposed, Disturbance of swallowing</td>
<td>Irritability, short tempered, Depressed, Stressed, Anxiety, Mood disturbances</td>
</tr>
</tbody>
</table>
Linking to the ICF (International Classification of Functioning)


<table>
<thead>
<tr>
<th>ICF category</th>
<th>QI-I</th>
<th>WHO</th>
<th>NHP</th>
<th>SF-36</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spitzer</td>
<td>DASII</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d450 Walking</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d4500 Walking short distances</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d4501 Walking long distances</td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>d455 Moving around</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>d4551 Climbing</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>d510 Washing oneself</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d530 Toileting</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d540 Dressing</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>d550 Eating</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d6309 Preparing meals, unspecified</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d640 Doing housework</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>d6509 Caring for household objects</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
8. Data synthesis – evaluating the evidence for the measurement properties

1. Homogeneity of the study characteristics
   Assess which studies are sufficiently similar to be synthesized

   • Homogeneity of study characteristics → study population, setting, (language) version of the instrument, mode of administration, design characteristics (time interval)

   • Methodological quality

   • Consistency of the results of the measurement properties
2. Data analysis

• Quantitative analysis
  Statistical pooling of the results is possible for reliability and correlation coefficients. To be recommended when studies are at least of fair quality.

• Qualitative analysis
  Best evidence synthesis: determine level of evidence taking into account the homogeneity of the studies, the methodological quality of the studies, and the consistency of the results.

3. Apply criteria of adequacy for the measurement properties
  Criteria are arbitrary – no consensus yet
Levels of evidence

<table>
<thead>
<tr>
<th>Level</th>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong</td>
<td>+++ or ---</td>
<td>Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality</td>
</tr>
<tr>
<td>moderate</td>
<td>++ or --</td>
<td>Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality</td>
</tr>
<tr>
<td>limited</td>
<td>+ or -</td>
<td>One study of fair methodological quality</td>
</tr>
<tr>
<td>conflicting</td>
<td>+/-</td>
<td>Conflicting findings</td>
</tr>
<tr>
<td>unknown</td>
<td>?</td>
<td>Only studies of poor methodological quality</td>
</tr>
</tbody>
</table>
Reliability of 6 of the 8 included questionnaires was assessed

NDI – 6 studies
NPDS – 2 studies
NBQ – 1 study
NPQ – 1 study
WDQ – 1 study
CNFDS – 1 study

Only 1 study of fair quality (NDI), other studies were of poor quality

Fair quality study found ICC=0.50

Conclusion “There is limited evidence that the reliability of the NDI is inadequate (ICC<0.70). The evidence for the reliability of the other questionnaires is unknown.”
Example content validity

For most questionnaires is was not clear what construct(s) the questionnaire intend to measure.

For none of the questionnaires, all aspects of content validity (relevance and comprehensiveness) were assessed.

For only 3 of the 8 questionnaires, it was assessed whether the items were relevant for the target population.

Results: Only part of the problems (NDI 33-55%, NDPS 64%, NPQ 44%) identified by patients, were included in the questionnaires.

Conclusion “Due to poorly described constructs of interest and poor quality of the studies, the evidence on the content validity of the included questionnaires is unknown”.

### Example

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Internal Consistency</th>
<th>Measurement Error</th>
<th>Reliability</th>
<th>Content Validity</th>
<th>Structural Validity</th>
<th>Hypothesis Testing</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDI</td>
<td>+++</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+/-</td>
<td>+++</td>
<td>+/-</td>
</tr>
<tr>
<td>NPDS</td>
<td>?</td>
<td>na</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>NBQ</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>na</td>
<td>na</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>WDQ</td>
<td>++</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>+</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>CNFDS</td>
<td>?</td>
<td>na</td>
<td>?</td>
<td>?</td>
<td>na</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CNQ</td>
<td>na</td>
<td>na</td>
<td>+</td>
<td>?</td>
<td>na</td>
<td>+</td>
<td>na</td>
</tr>
<tr>
<td>CWOM</td>
<td>?</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>+</td>
</tr>
</tbody>
</table>
9. Drawing a conclusion

• The key elements of the research question should be reflected in the conclusions of the review.

• To decide upon the best instrument, all measurement properties should be considered together.

• Take into account the number of studies, the methodological quality, and the consistency of the results.

• Mention the methods and criteria used in the data-synthesis process.
“For clinical practice and research in patients with neck pain we recommend using the NDI to measure pain and disability, because it is the questionnaire for which the most information on its measurement properties is available and the results are mostly positive. However, research is needed to clarify its underlying constructs, reliability, and responsiveness in (subgroups of) neck pain patients.

This does not mean that the other questionnaires are poor, but implies that studies of high methodological quality are needed to properly assess their measurement properties.”
The publication about the review should contain:

1. the results of the search and study selection
2. the methodological quality of the included studies
3. the characteristics of the measurement instruments
4. the characteristics of the included study population
5. the results of the measurement properties
6. the conclusion about the best measurement instrument
Exercise

Data synthesis:

combining the results from different studies
Construct: disability

Based on conceptual model of the WHO

WHO’s definition of disability: “any restriction or lack of ability to perform an activity in a manner or within the range considered normal for a human being”.

Disability was operationalized in terms of difficulty experienced while performing simple tasks
Item selection: literature, patients, experts

Item reduction: 48 items were reduced based on factor analysis and IRT methods to 20 items

Format: 20 items, regarding 6 domains (bed/rest, sitting/standing, ambulation, movement, bending/stooping, handling large/heavy objects)

Items scored from 0-5, and summarized in one score from 0-100 (100=maximum disability)
Today, do you find it difficult to perform the following activities because of your back?
1. Get out of bed.
2. Sleep through the night.
3. Turn over in bed.
4. Ride in a car.
5. Stand up for 20–30 minutes.
6. Sit in a chair for several hours.
7. Climb one flight of stairs.
8. Walk a few blocks (300–400 m).
9. Walk several miles.
10. Reach up to high shelves.
11. Throw a ball.
12. Run one block (about 100 m).
13. Take food out of the refrigerator.
14. Make your bed.
15. Put on socks (pantyhose).
16. Bend over to clean the bathtub.
17. Move a chair.
18. Pull or push heavy doors.
19. Carry two bags of groceries.
20. Lift and carry a heavy suitcase.

Response options
0 not difficult at all
1 minimally difficult
2 somewhat difficult
3 fairly difficult
4 very difficult
5 unable to do
Exercise 1: data synthesis reliability

9 studies

2. Rodrigues et al. Spine 2009
6. Davidson et al. Phys Ther 2002
1. **Homogeneity of the study characteristics**
   Assess which studies are sufficiently similar to be synthesized.

2. **Quantitative analysis**
   Statistical pooling of the results possible for reliability and correlation coefficients. To be recommended when studies are at least of moderate quality.

3. **Qualitative analysis**
   Best evidence synthesis: determine level of evidence taking into account the methodological quality of the studies, the consistency of the results, and the homogeneity of the studies.

4. **Adequacy of the measurement properties**
   Criteria are arbitrary – no consensus yet.
Step 1

Assess which studies are sufficiently similar to be synthesized. Consider:

1. Study characteristics → Table 1

2. Methodological quality → Table 2 (and COSMIN box Reliability)

3. Results of the measurement properties → Table 3
Discuss:

How can the results be combined:

Statistical pooling or best evidence synthesis?
Step 3

Perform a best evidence synthesis. Consider:

1. Levels of evidence → Table 4
2. Criteria of adequacy → Table 5
Points of discussion

1. What kind of studies do you not want to combine?
   - results from different language versions?
   - results from patients with different disease characteristics?

2. Should poor quality studies be excluded from the data synthesis?

3. Should studies with inconsistent results not be combined?
4. Should we try to statistically pool the results of the measurement properties (e.g. reliability coefficients)?
Exercise 2: data synthesis construct validity

10 studies

2. Rodrigues et al. Spine 2009

Step 1

Assess which studies are sufficiently similar to be synthesized. Consider:

1. Study characteristics → Table 1

2. Methodological quality → Table 2 (and COSMIN box Construct validity (hypotheses testing))

3. Results of the measurement properties → Table 3
Discuss:

How can the results be combined:

Statistical pooling or best evidence synthesis?
Step 3

Perform a best evidence synthesis. Consider:

1. Levels of evidence → Table 4
2. Criteria of adequacy → Table 5
Summary – 10 steps

1. formulating a research question
2. performing a literature search
3. formulating eligibility criteria
4. selecting abstracts and full-text articles
5. evaluating the methodological quality of the included studies
6. extracting the data
7. comparing the content
8. data synthesis- evaluating the evidence for adequacy of the measurement properties
9. drawing an overall conclusion of the systematic review
10. reporting on the systematic review
Why a systematic review of measurement properties?

- Systematic reviews of measurement properties are fun to do
  
  Good learning experience for students
  
  Good opportunity for cooperation with other people
  
  Reviews are cited a lot
Knowledgecenter Measurement Instruments: [www.kmin-vumc.nl](http://www.kmin-vumc.nl)

COSMIN: [www.cosmin.nl](http://www.cosmin.nl)

Cochrane PRO methods group: [www.cochrane-pro-mg.org](http://www.cochrane-pro-mg.org)

E-mail: [hcw.devet@vumc.nl](mailto:hcw.devet@vumc.nl)  
[cb.terwee@vumc.nl](mailto:cb.terwee@vumc.nl)  
[w.mokkink@vumc.nl](mailto:w.mokkink@vumc.nl)  
[donald@u.washington.edu](mailto:donald@u.washington.edu)