# Protocol for systematic reviews of measurement properties

November 2011

Caroline B Terwee, PhD
VU University Medical Center
Knowledgecenter Measurement Instruments
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
P.O.Box 5057
1007 MB Amsterdam
E-mail: cb.terwee@vumc.nl

COSMIN: www.cosmin.nl

## Introduction

A systematic review of measurement properties is a useful tool to get a clear and comprehensive overview of the measurement properties of (all) measurement instruments, and to come to a conclusion about the best instrument available for a particular purpose. In this document the methodology of systematic reviews of measurement properties is exaplined, based on ten steps, as described in Table 1 [1]. It should be noted, however, that part of the methodology is still under development.

**Table 1. Ten steps to conduct a systematic review of measurement properties**

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. data extraction
7. content comparison
8. data syntheses - evaluating of quality of the instruments
9. overall conclusion of the systematic review
10. reporting a systematic review of measurement properties

**Step 1. Research question**

In the research question four key elements should be included:

(1) the construct of interest or the name(s) of the measurement instrument(s) of interest

(2) the population of interest

(3) the type of measurement instrument of interest, e.g. imaging techniques, laboratory tests, observation scales, performance based instruments, interviews or questionnaires, etc.

(4) the measurement properties on which the review focuses.

Example: The aim of this study is to critically appraise the evidence on the measurement properties of questionnaires measuring functioning in patients with hand osteoarthritis (OA).

**Step 2. Literature search**

We recommend using at least MEDLINE (e.g. using the PubMed interface) and EMBASE. In addition, databases focusing on particular professional organisations may be searched, such as PsycINFO or CINAHL.

A search strategy consist of collections of search terms for the following characteristics

1) construct of interest
2) target population
3) measurement properties

For each of these searches a comprehensive list of possible synonyms should be made, consisting of index terms (e.g. 'MESH terms' (Medical Subject Headings) in MEDLINE or EMTREE terms in EMBASE), combined with free text words. These synonyms should be combined with the conjunction 'OR'. The searches for these three characteristics should then be combined with the conjunction 'AND', to obtain the list of references that should be used to select the relevant articles. Selecting adequate search terms and building the search strategy should be done by an expert on the specific construct in close cooperation with a medical information specialist.

We advise against the use of search terms indicating the type of measurement instruments because many studies don't use specific terms and there is a high risk of missing relevant articles when search terms for type of instrument are used. Exception is the search for patient-reported outcomes in Pubmed. For this purpose, a instrument filter can be used for finding studies on patient-reported outcomes, developed by the PROM group, University of Oxford (contact Elizabeth Gibbons elizabeth.gibbons@dph.ox.ac.uk or Caroline Terwee cb.terwee@vumc.nl).

In addition to the search strategy described above, we recommend performing an additional search including the names of the instruments which are found in the initial search. These names can be combined, using the AND conjunction, with terms for the target population and measurement properties. Furthermore, we recommend checking the references of the articles included in the review to search for additional relevant studies.

We recommend not using a time limit in the search because older literature on measurement properties is still relevant. Language restrictions are also not recommended, but for practical reasons the review is often restricted to articles written in those languages that are well mastered by the researchers.

Finally, we recommend to consult a clinical librarian for fine-tuning the searches.

Example:
The following databases will be searched for a systematic review on questionnaires measuring functioning in patients with hand osteoarthritis (OA): PubMed (1966-2010), Embase (1974-2010), CINAHL (EBSCOhost) (1981-2010), and PsycINFO (EBSCOhost) (1806-2010). The search will contain blocks of search terms related to the following aspects:
(1) construct of interest (functioning): no search terms for functioning will be used. Instead, questionnaires measuring functioning will be selected by hand from the search;
(2) target population (hand OA): different terms for hand OA will be used, such as ''osteoarthritis'' and ''arthritis'', combined with ''hand'';
(3) type of instrument (questionnaire); In Pubmed, an instrument filter will be used for finding studies on patient-reported outcomes, developed by the PROM group, University of Oxford (personal communication). In the other databases no search terms will be used;
(4) measurement properties: In Pubmed a validated search filter for studies on measurement properties will be used [2]. In the other databases, a selection of relevant search terms will be used, which has been used in previous reviews.
An additional search will be performed in each database, including the names of the instruments which are found in the initial search. Reference lists will be screened to identify additional relevant studies.

An example of a full search strategy is provided in Appendix 1

**Step 3. Eligibility criteria**

We recommend using at least the following inclusion criteria (again using the four key elements as in the research question):

(1) the instruments should aim to measure the construct of interest.
(2) the study sample should concern the target population of interest
(3) the study should concern the type of measurement instrument of interest
(4) the aim of the study should be the development of a measurement instrument or the evaluation of one or more of its measurement properties. Studies that only focus on interpretability, e.g. the determination of minimal important change, can also be included.
(5) The study should be published as a full text original article.

Often, much indirect evidence on measurement properties of instruments can be obtained, e.g. from studies in which the instrument of interest is used in the validation process of another instrument, or in an RCT or other longitudinal study in which indirect evidence for responsiveness might be found. We recommend excluding these kinds of studies from the review for two reasons. First of all, it is very difficult to find all of these articles in a manageable and structured way. Secondly, it is often difficult to interpret the evidence for validity or responsiveness provided in these studies, because no hypotheses about the validity or responsiveness of the instrument of interest are formulated and tested in these studies.

Example:

- The questionnaire should aim to measure functioning (according to the developers of the instrument). Functioning is defined accoring to the International Classification of Functioning as Activity Limitations, which are difficulties an individual may have in executing activities.
- The study population should be adults (age ≥18) with hand OA, which is defined as radiologically verified OA with Kellgren and Lawrence grade ≥2 in any hand joint.
- The instrument under study should be a self-report questionnaire.
- The aim of the study should be the development of a measurement instrument or the evaluation of one or more of its measurement properties. Studies that only focus on interpretability, e.g. the determination of minimal important change, will also be included.
- The study should be published as a full text original article.
- Articles in all languages will be included.

Exclusion critiria are:

- Studies in patients with other hand conditions, e.g., carpal tunnel syndrome or other types of arthritis such as rheumatoid arthritis (RA).
- Trials or studies evaluating the effectiveness of interventions where a questionnaire is used as an endpoint (without studying the measurement properties).
- Questionnaires administered by interview or proxy.

**Step 4. Selection of abstracts and full-text articles**

Two reviewers should independently assess titles, abstracts, selected full-text articles, and reference lists of the studies retrieved by the literature search. In case of disagreement between the two reviewers, a third reviewer will make the decision regarding inclusion of the article.

The search should be carefully documented. The names of the databases that were searched, as well as the interface that was used to search the databases, such as PubMed or OVID for searching Medline, should be documented. Also the date of the search, the exact search terms, and any limitations (e.g. language or age restrictions) that were used, should be documented. Next, we recommend to carefully document the titles initially selected, the full-text articles retrieved, and the articles included in the review. The reasons for exclusion of retrieved full text articles are also useful to document. We recommend presenting all this information on searching and selection in a flow chart.

**Step 5. Evaluation of the methodological quality of the included studies**

The methodological quality of the included studies should be assessed using the COSMIN checklist 13. This checklist consists of nine boxes with standards for how each measurement property should be assessed. Each item will be scored on a 4-point rating scale (i.e. “poor”, “fair”, “good” or “excellent”), which is an additional feature of the COSMIN checklist (www.cosmin.nl) [3-5]. The COSMIN taxonomy and definitions will be used to decide which measurement properties have been evaluated in a study and which corresponding boxes should be completed. An overall score for the methodological quality of a study is determined by taking the lowest rating of any of the items in a box. For studies that used Item Response Theory analyses, the COSMIN IRT box with questions on e.g. software and method of estimation, will also be completed and taken into account in the quality rating for the measurement properties assessed with the IRT analyses.

Assessment of the methodological quality should be performed by two reviewers independently. In case of disagreement, a third reviewer will make the decision.

In systematic reviews of measurement properties the Interpretability box and the Generalizability box are mainly used as data extraction forms (see step 6). Therefore, these boxes do not need to be completed as part of the methodological quality assessment.

**Step 6. Data extraction**

We recommend the following information to be extracted from the included articles by two reviewers independently:

- General characteristics of the instruments (construct, subscales, # items, version, etc)
- Characteristics of the study populations in which the measurement properties were assessed (age, gender, disease severity, setting, country, language). The information mentioned in items 1-6 from the COSMIN box Generalisability is extracted from the studies.
- Results of the measurement properties.
- Evidence on the interpretability of the included questionnaires. For each study, the information described in items 4-8 of the COSMIN box Interpretability (information on distribution of scores, floor- and ceiling effect, and minimal important change) will be extracted.

A data extraction form can be developed for this purpose or the data can be directly extracted into Tables.

**Step 7. Content comparison**

A content comparison is an overview of the content of each instrument, on item level, which is useful to examine which content is covered by the different instruments. For example, the content of all measurement instruments assessing functional or health status can be linked to the International Classification of Functioning, Disability and Health (ICF). Such a content comparison has for example been made by Stamm et al. in 2006 of questionnaires measuring functioning in patients with hand OA [6].

**Step 8. Data synthesis – evaluating the quality of the measurement instruments**

The next step is to evaluate the quality of the instruments itself, i.e. to evaluate the measurement properties of the included instruments. To determine the overall evidence for the measurement properties of the instruments the results of the different studies, adjusted for their methodological quality, will be combined.

Combining results of different studies on a measurment property of an instrument is only possible when the studies are sufficiently similar with regard to study population and setting, the (language) version of the instrument that is used, and the form of administration. To

judge the similarities of different study populations, the data extracted with the Generalisability box can be used.

Because pooling of measurement properties is still under development, often a best evidence synthesis is performed. The possible overall rating for a measurement property is “positive”, “indeterminate”, or “negative”, accompanied with a level of evidence (strong, moderate, limited, conflicting, unknown), as proposed by the Cochrane Back Review Group (Table 1) 18,19. To give a positive or negative rating for the results of the measurement properties, criteria for good measurement properties will be used, based on criteria proposed by Terwee et al. [7] (Table 2).

For some measurement properties, rating the results and applying levels of evidence is more complex because information from different studies should be taken into account. These situations are described below:

- Internal consistency: for a proper judgment, information is needed on unidimensionality of the scales as well as on Cronbach’s alpha. This information may come from different studies. For rating strong evidence for a positive internal consistency, there should be consistent findings in at least two studies of good methodological quality or one study of excellent methodological quality that the (sub)scales are unidimensional. In addition, there should be consistent findings in at least two studies of good quality or one study of excellent quality that the Cronbach’s alpha is at least 0.70.
- Reliability: Levels of evidence, as described in Table 2, can be applied. The evidence will be graded as strong when consistent positive results (ICCs or Kappas >0.70) are found in at least two studies of good quality or one study of excellent quality.
- Measurement error: here information is needed on the Smallest Detectable Change (SDC) as well as on the Minimal Important Change (MIC). Again, this information may come from different studies. For strong evidence, the SDC should be calculated (or possible to deduce from the reported data in the article) in at least two studies of good quality or one study of excellent quality. In addition, the MIC should be calculated in at least two studies of good quality or one study of excellent quality. Next, an estimation should be made whether the SDC is smaller than the MIC. Ideally, this should be based on comparing the pooled estimate of the SDC with the pooled estimite of the MIC, but the methodology of pooling SDCs or MICs is still under development. Therefore, a subjective judgement will be made.
- Content validity: Different aspects of content validity (see COSMIN box D) can be evaluated in different studies. Strong evidence for a positive content validity will be rated if all four aspects have adequately been evaluated with positive results, i.e all items are considered relevant for the construct, purpose, and target population and the instrument is considered comprehensive. One study per aspect is considered sufficient evidence.

Moderate evidence will be rated if two or three aspects have been evaluated, i.e. the items are considered relevant for the construct or target population (items 1 or 2) and the instrument is considered comprehensive (item 4). Limited evidence is rated when only one aspect of content validity is assessed.

- Structural validity: Levels of evidence, as described in Table 2, will be applied. For rating strong evidence for a positive structural validity, there should be consistent findings in at least two studies of good methodological quality or one study of excellent methodological quality that the expected factor structure is confirmed.
- Construct validity and responsiveness (hypotheses testing): validation is an ongoing process and therefore no strict criteria can be defined for good construct validity or responsiveness. The total evidence from all included studies will be judged together, based on the criteria described in Table 2.
- Criterion validity: Levels of evidence, as described in Table 2, will be applied. The evidence will be graded as strong when consistent positive results (correlations with gold standard >0.70) are found in at least two studies of good quality or one study of excellent quality.
- Cross-cultural validity: Levels of evidence, as described in Table 2, will be applied. For rating strong evidence for a positive cross-cultural validity, there should be consistent findings in at least two studies of good methodological quality or one study of excellent methodological quality that the expected factor structure is confirmed for the new language version and that no important DIF between language versions has been found.

Note that this step is still under development.

**Step 9. Overall conclusion**

To select the *best* measurement instrument for a particular situation, the number of studies in which the measurement properties of the instrument is investigated, the methodological quality of these studies, and (the consistency of) the results of those studies should be taken into account. Conclusions should be drawn over studies with sufficient clinical homogeneity, i.e. similarities with regard to construct measured, purpose and study population. Thus, the conclusion of a systematic review may be that a measurement instrument is (the most) appropriate to measure a construct in a specific population, but not in another population.

**Step 10. Reporting**

The results of the review will be reported, following the PRISMA guidelines (www.prisma-statement.org). We recommend the following things to be reported:

- the results of the literature search and selection of the studies (presented in a flow chart)

- the methodological quality of the included studies (table or graph)
- the characteristics of the included questionnaires (table)
- the characteristics of the included study populations (table)
- the results of the measurement properties (table)
- Levels of evidence for all measurement properties of the included questionnaires (table)
- the conclusion about the best measurement instrument (in the text) for a given population or purpose.

**Final remark**

A systematic review on measurement properties actually consists of a collection of separate systematic reviews per measurement property. That these are separate reviews becomes visible in the number of studies that contribute data to each measurement property, a separate set of items (COSMIN box) per measurement property to appraise the methodological quality of the studies, and separate methods of data synthesis per measurement property. However, to draw a conclusion about the choice of the best measurement instrument in a particular situation, the results of multiple measurement properties should be taken into account. Therefore a systematic review usually contains information on all measurement properties. This means that conducting such a review can be quite complex and time consuming. However, if much evidence is available it is better to write an informative review on a few measurement instruments, or separate reviews for each measurement property than to write a superficial mega-review which lacks much relevant information.

## References


[1] de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge: Cambridge University Press; 2011.

[2] Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115-23.

[3] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010;19:539-49.

[4] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study. Journal of Clinical Epidemiology 2010;63:737-45.

[5] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodeology 2010;10:22.

[6] Stamm T, Geyh S, Cieza A, Machold K, Kollerits B, Kloppenburg M, et al. Measuring functioning in patients with hand osteoarthritis--content comparison of questionnaires based on the International Classification of Functioning, Disability and Health (ICF). Rheumatology (Oxford) 2006;45:1534-41.

[7] Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.

**Table 1 – Levels of evidence for the quality of the measurement property**

| Level | Rating† | Criteria |
|---|---|---|
| strong | **+++ or ---** | Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality |
| moderate | **++ or --** | Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality |
| limited | **+ or -** | One study of fair methodological quality |
| conflicting | **+/-** | Conflicting findings |
| unknown | **?** | Only studies of poor methodological quality |

(..) = reference number,

† + = positive rating, ? = indeterminate rating, - = negative rating

**Table 2 – Quality criteria for measurement properties [7]**

| Property | Rating† | Quality Criteria |
|---|---|---|
| **Reliability** | | |
| Internal consistency | + | Cronbach's alpha(s) ≥ 0.70 |
| | ? | Cronbach's alpha not determined or dimensionality unknown |
| | - | Cronbach's alpha(s) < 0.70 |
| Reliability | + | ICC / weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80 |
| | ? | Neither ICC / weighted Kappa, nor Pearson's r determined |
| | - | ICC / weighted Kappa < 0.70 OR Pearson's r < 0.80 |
| Measurement error | + | MIC > SDC OR MIC outside the LOA |
| | ? | MIC not defined |
| | - | MIC ≤ SDC OR MIC equals or inside LOA |
| **Validity** | | |
| Content validity | + | All items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement AND the questionnaire is considered to be comprehensive |
| | ? | Not enough information available |
| | - | Not all items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement OR the questionnaire is considered not to be comprehensive |
| Construct validity - Structural validity | + | Factors should explain at least 50% of the variance |
| | ? | Explained variance not mentioned |
| | - | Factors explain < 50% of the variance |
| - Hypothesis testing | + | Correlations with instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses AND correlations with related constructs are higher than with unrelated constructs |
| | ? | Solely correlations determined with unrelated constructs |
| | - | Correlations with instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR correlations with related constructs are lower than with unrelated constructs |
| - Cross-cultural validity | + | No differences in factor structure OR no important DIF between language versions |
| | ? | Multiple group factor analysis not applied AND DIF not assessed |
| | - | Differences in factor structure OR important DIF between language versions |
| Criterion validity | + | Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥ 0.70 |
| | ? | No convincing arguments that gold standard is "gold" |
| | - | Correlation with gold standard < 0.70 |
| **Responsiveness** | | |
| Responsiveness | + | Correlation with changes on instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70 AND correlations with changes in related constructs are higher than with unrelated constructs |
| | ? | Solely correlations determined with unrelated constructs |
| | - | Correlations with changes on instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 OR correlations with changes in related constructs are lower than with unrelated constructs |

MIC = minimal important change, SDC = smallest detectable change, LoA = limits of agreement, ICC = intraclass correlation coefficient, DIF = differential item functioning, AUC = area under the curve

† + = positive rating, ? = indeterminate rating, - = negative rating

## Appendix 1 Example search strategy

This search strategy was used in a systematic review on the measurement properties of studies on questionnaires measuring functioning in patients with hand OA.

**MEDLINE (PubMed)**

(Arthritis[MeSH] OR Arthritis[tiab] OR Osteoarthritis[MeSH] Osteoarthritis [tiab]) AND hand[tiab])

**AND**

(instrumentation[sh] OR methods[sh] OR "Validation Studies"[pt] OR "Comparative Study"[pt] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR "outcome assessment"[tiab] OR "outcome measure*"[tw] OR "observer variation"[MeSH] OR "observer variation"[tiab] OR "Health Status Indicators"[Mesh] OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR "coefficient of variation"[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tw] OR precision[tw] OR imprecision[tw] OR "precise values"[tw] OR test-retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] OR result[tw] OR results[tw] OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR "factor analysis"[tiab] OR "factor analyses"[tiab] OR "factor structure"[tiab] OR "factor structures"[tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR "item discriminant"[tiab] OR "interscale correlation*"[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR "interval variability"[tiab] OR "rate variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR "minimal detectable concentration"[tiab] OR interpretab*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab])

**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]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

**Embase (www.embase.com)**

(Arthritis OR Osteoarthritis) AND hand)
**AND**
('questionnaire'/exp OR 'named inventories, questionnaires and rating scales'/exp OR 'psychometry'/exp OR 'outcome assessment'/exp OR 'pain assessment'/exp OR 'disability'/exp OR 'validity'/exp OR 'reliability'/exp)

**CINAHL (EBSCOhost)**

(Arthritis OR Osteoarthritis) AND hand)
**AND**
((MH "Research Measurement+") OR (MH "Outcome Assessment") OR (MH "Outcomes Research")))

**PsycINFO (EBSCOhost)**

(Arthritis OR Osteoarthritis) AND hand)
**AND**
AND (exp measurement/ OR exp test construction/ OR exp interrater reliability/ OR exp statistical analysis/)