


BMJ Open Tools to assess the measurement properties of quality of life instruments: a meta-review

Sonia Lorente ,^{1,2} Carme Viladrich,¹ Jaume Vives,^{1,3} Josep-Maria Losilla^{1,3}

To cite: Lorente S, Viladrich C, Vives J, *et al.* Tools to assess the measurement properties of quality of life instruments: a meta-review. *BMJ Open* 2020;**10**:e036038. doi:10.1136/bmjopen-2019-036038

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-036038>).

Received 27 November 2019
Revised 15 June 2020
Accepted 06 July 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Psychobiology and Methodology of Health Science. Area of Behavioral Science Methodology, Universitat Autònoma de Barcelona, Cerdanyola del Vallés, Barcelona, Spain

²Pediatric Area, Consorci Sanitari de Terrassa, Terrassa, Spain

³Sport Research Institute, Universitat Autònoma de Barcelona, Cerdanyola, Spain

Correspondence to

Dr Jaume Vives;
Jaume.Vives@uab.cat

ABSTRACT

Objective This meta-review aims to discuss the methodological, research and practical applications of tools that assess the measurement properties of instruments evaluating health-related quality of life (HRQoL) that have been reported in systematic reviews.
Design Meta-review.

Methods Electronic search from January 2008 to May 2020 was carried out on PubMed, CINAHL, PsycINFO, SCOPUS, WoS, Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) database, Google Scholar and ProQuest Dissertations and Theses.

Results A total of 246 systematic reviews were assessed. Concerning the quality of the review process, some methodological shortcomings were found, such as poor compliance with reporting or methodological guidelines. Regarding the procedures to assess the quality of measurement properties, 164 (66.6%) of reviewers applied one tool at least. Tool format and structure differed across standards or scientific traditions (ie, psychology, medicine and economics), but most assess both measurement properties and the usability of instruments. As far as the results and conclusions of systematic reviews are concerned, only 68 (27.5%) linked the intended use of the instrument to specific measurement properties (eg, evaluative use to responsiveness).

Conclusions The reporting and methodological quality of reviews have increased over time, but there is still room for improvement regarding adherence to guidelines. The COSMIN would be the most widespread and comprehensive tool to assess both the risk of bias of primary studies, and the measurement properties of HRQoL instruments for evaluative purposes. Our analysis of other assessment tools and measurement standards can serve as a starting point for future lines of work on the COSMIN tool, such as considering a more comprehensive evaluation of feasibility, including burden and fairness; expanding its scope for measurement instruments with a different use than evaluative; and improving its assessment of the risk of bias of primary studies.

PROSPERO registration number CRD42017065232.

INTRODUCTION

The systematic reviews of measurement properties critically appraise the content and measurement properties of all instruments that assess a certain construct of interest in

Strengths and limitations of this study

- The search strategy has been designed to be comprehensive, following the Peer Review of Electronic Search Strategies guidelines including specific filters for finding studies on psychometric properties of measurement instruments.
- A total of 246 systematic reviews were included and, to our knowledge, this meta-review provides the broadest overview of the most common tools used to assess measurement properties of health-related quality of life instruments and their relationship with measurement standards, scientific traditions and the intended use of the measures.
- Some of the included systematic reviews poorly reported the review process, outcomes and conclusions, and this fact may have led to the loss of some data.
- Inclusion of studies published in English only may have led to language bias.

a specific study population.¹ These systematic reviews provide both a comprehensive overview of the measurement properties of health instruments and supportive evidence for the selection of instruments for a specific purpose (eg, research, clinical practice, predictive).^{2–3} In this type of systematic review, different authors have evaluated not only the methodological quality of their key phases—namely the search strategy, the bias risk assessment of the primary studies and the data synthesis—but also whether the measurement properties of the health status instruments have been appraised with standardised procedures or tools during the data extraction phase.^{1 2 4 5} However, depending on the measurement standards on which these tools were developed, the approach to analyse the measurement properties of instruments may vary.⁶ This could lead to different conclusions and recommendations, in spite of the effort undertaken by the international Society for Quality of Life Research to set consensus-based minimum standards.⁷ Besides, according to Rosenkoetter and Tate,⁶

the assessment tools commonly used by clinicians and researchers to select the appropriate outcome measures for specific purposes show a variety of forms and cover a mix of standards related to reporting, methodological quality and statistical outcome quality.

The aims of this present meta-review are to: (1) identify systematic reviews assessing the measurement properties of health-related quality of life (HRQoL) instruments; (2) identify the main tools applied to assess their measurement properties; (3) describe the contents of the applied tools (validity, reliability, feasibility, etc); (4) identify the measurement standards on which these tools were developed or conform to, comparing their similarities and differences and (5) appraise how authors of these systematic reviews include the assessment of the measurement quality in their results and conclusions, that is, to what extent conclusions depend on the results of the evaluation of the measurement properties, as well as their relationship, if any, with the intended use of the HRQoL instrument (eg, evaluative).

METHODS

The protocol of this review⁸ was prospectively registered. We conducted this meta-review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA).^{9 10}

Search strategy

A systematic search was performed in PubMed, US National Library of Medicine, by National Center for Biotechnology Information (NCBI); CINAHL, Cumulative Index to Nursing and Allied Health Literature, by EBSCOhost; PsycINFO, Psychological Information, by APA PsycNET; SCOPUS by Elsevier; WoS, Web of Science CORE, by Thomson Reuters; Consensus-based Standards for the selection of Health Measurement Instruments database, by COSMIN Initiative (<http://www.cosmin.nl/>); and Google Scholar (up to 400 links). ProQuest Dissertations and Theses Global was used for searching grey literature, and search alerts in all databases were set. The search strategy followed the Peer Review of Electronic Search Strategies guidelines recommendations,^{11 12} and consisted of three filters composed of search terms for the following: (1) systematic review methodology; (2) HRQoL instruments and (3) measurement properties. The latter filter was developed by the Vrije University Medical Center for finding studies on measurement properties of measurement instruments.¹³ All filters were adapted for all databases. The searches were completed in May 2020. Restrictions by language (English) and publication date (from January 2008) were applied (see online supplementary file 1 for search strings for all databases).

Inclusion criteria

Systematic reviews specifically aiming to report or to assess the measurement properties of instruments evaluating the quality of life within the context of health and

disease¹⁴ were included. Systematic reviews were required to include the full results report, and detailed information about the procedures used to assess the measurement properties.

Exclusion criteria

Systematic reviews exclusively focused on evaluating clinical interventions were excluded. Systematic reviews specifically focused on assessing patient-reported outcomes measures (PROMs) other than HRQoL for specific diseases, clinical conditions or populations, were excluded. Systematic reviews that did not report full information about the procedures to assess the measurement properties were also excluded (eg, conference abstracts).

Study screening

References identified by the search strategy were entered to Mendeley reference management software, and duplicates were removed. Titles and abstracts were screened independently by two reviewers (SL and JV). When decisions were unable to be made from title and abstract alone, the full paper was retrieved. Full-text inclusion criteria were checked independently by two reviewers (SL and JV). Discrepancies during the process were resolved through discussion (with independent reviews of J-ML and CV when necessary).

Data extraction

Extracted information of each selected systematic review and meta-analysis included general information such as author, year and quality of review process of systematic reviews (eg, protocol registration, reporting guidelines and use of flow chart). Information concerning the main identified tools applied to assess the measurement properties of HRQoL instruments included the title, intended use, number of items, response categories, instrument assessment criteria and measurement properties assessed. Information on how authors included the assessment of the quality of HRQoL in their results and conclusions was also extracted. Authors of eligible studies were contacted to provide missing or additional data when necessary.

Study aim

To examine the methodological, research and practical applications of the reported tools in systematic reviews that assess the measurement properties of instruments evaluating quality of life within the context of health and disease, that is, HRQoL.

RESULTS

Search results

Figure 1 shows the results of the search strategy, reported according to the PRISMA flow diagram. A total of 4320 references were identified through database searches. After removing duplicates, 3055 titles and abstracts were screened. After the assessment of 525 full-text documents for eligibility, a total of 246 systematic reviews were included in the qualitative analysis. These systematic

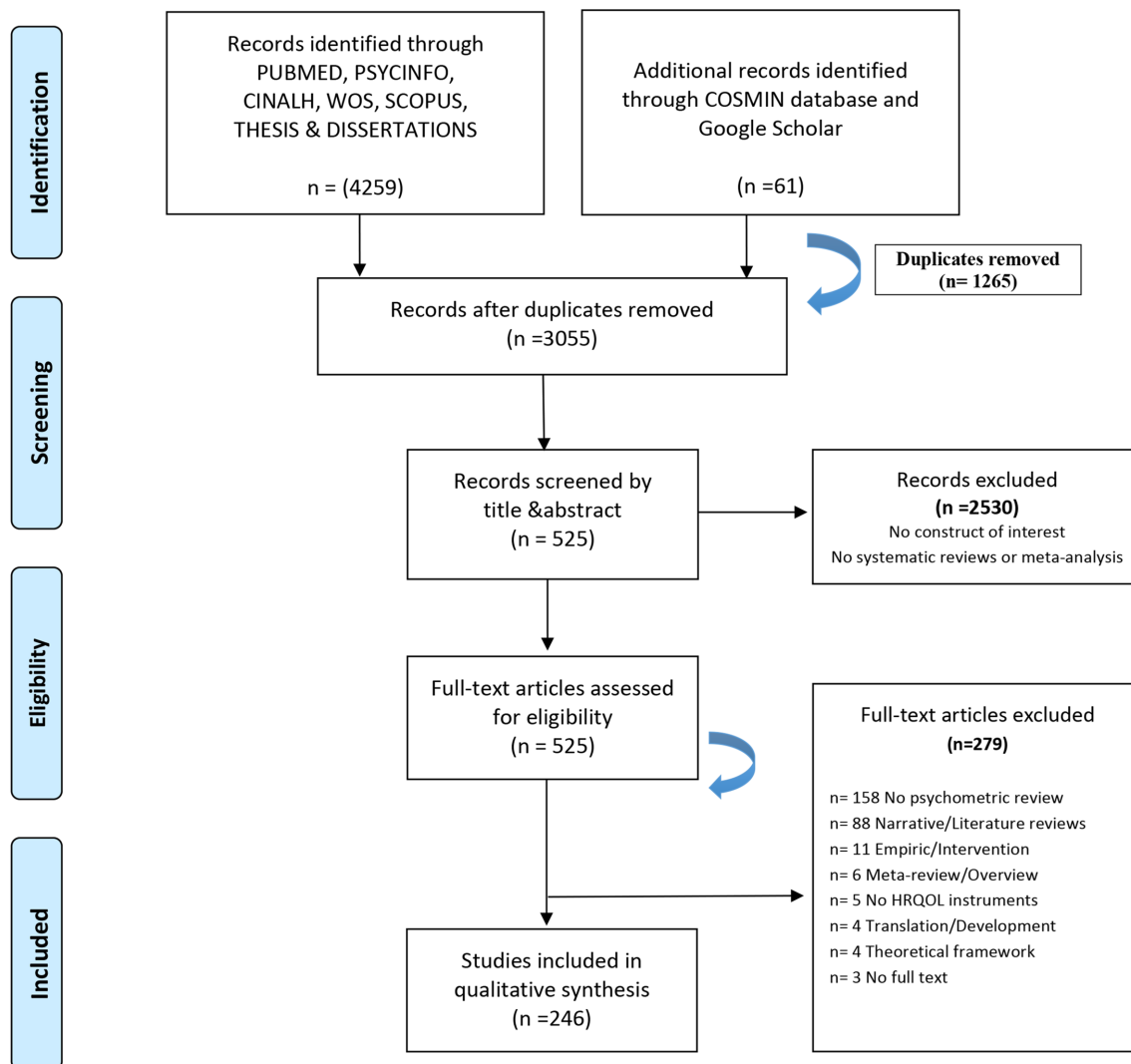


Figure 1 PRISMA flow chart. Flow diagram for search results (from Moher *et al*⁹). COSMIN, Consensus-based Standards for the selection of health Measurement Instruments; HRQoL, health-related quality of life; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

reviews covered a wide range of HRQoL instruments, both generic and disease specific. A total of 24 (9.8 %) of the systematic reviews assessed the quality of one measurement property only, such as the conceptual and measurement model or the content validity (see online supplementary file S2 for characteristics and references of studies).

Reporting and methodological quality of the studies

Table 1 shows the reporting and methodological quality of systematic reviews. Findings showed that 27 (10.9%) of the reports registered the protocol prospectively, a figure that raised to 20.8% when considering the reports from 2014 onwards; 78 (31.7%) followed reporting guidelines such as PRISMA (50.8% the last 6 years); 42 (17.0% since 2008; 23.8% for the last 6 years) assessed the reporting and/or the methodological quality of primary studies using recommended guides, such as Standards for the Reporting of Diagnostic Accuracy Studies and Quality Assessment of Diagnostic Accuracy Studies, respectively;

238 (96.7 %) reported the search strategy; 116 (47.41%) reported the detailed syntax for one database at least; 134 (54.4%) made the article selection by two or more independent reviewers; 166 (67.5%) used a flow chart to report search outcomes and 132 (53.7%) stated the funding. These last percentages slightly increased when reducing the time frame to the last 6 years.

Assessment of measurement properties of HRQoL instruments

Assessment procedures of measurement properties varied considerably. A total of 164 (66.6%) out of 246 systematic reviews applied one tool at least, that is, a published and well-accepted list of criteria, to rate the evidence on measurement properties of instruments; 41 (16.6%) applied their own author's criteria only; 30 (12.2%) followed literature recommendations included in very highly circulated books or papers only, and 14 (5.7%) used an ad hoc checklist of criteria only. A total of 98 (39.8%) systematic reviews did combine different

Table 1 Reporting and methodological quality of studies

	2008–2020		2014–2020	
	N	%	N	%
Protocol registered prospectively				
▶ Yes, PROSPERO	27	10.9	26	20.5
▶ No registered	219	89.1	100	79.3
Standards of systematic review reporting and/or quality assessment				
▶ Yes (AMSTAR, PRISMA, QUOROM...)	78	31.7	64	50.8
▶ No	168	68.3	62	49.2
Standards to assess reporting and/or quality assessment of primary studies				
▶ Yes (QUADAS, STARD...)	42	17.0	30	23.8
▶ No	204	83.0	96	76.2
No of databases searched				
▶ 1–3	96	39.1	50	39.6
▶ 4–6	107	43.4	61	48.4
▶ 7–9	22	8.9	8	6.3
▶ ≥10	18	7.3	6	4.7
▶ Not reported	3	1.2	1	0.8
Other sources				
▶ Official websites/internet	25	10.1	7	5.5
▶ Virtual libraries	24	9.7	12	9.4
▶ Google/google scholar	25	10.1	14	11.0
▶ Scientific journals/thesis	6	2.4	2	1.6
Search strategy				
Terms, databases, time period				
▶ Yes	238	96.7	123	97.6
▶ No	8	3.3	3	2.4
Search syntax				
▶ Detailed syntax reported (Truncations, Booleans...)	115	46.7	79	62.7
▶ Syntax not reported or not detailed enough to be replicable	125	50.8	46	36.5
▶ Supplementary file under request (not available)	5	2.1	1	0.8
Inclusion/exclusion selection criteria				
▶ Reported and well-defined	229	93.1	122	96.8
▶ Not reported or not clearly stated	17	6.9	4	3.2
Article selection				
▶ By two or more independent reviewers	134	54.4	87	69.0
▶ Not reported or not clearly stated	112	45.6	39	31.0
Flow chart				
▶ Yes	166	67.5	108	85.7

Continued

Table 1 Continued

	2008–2020		2014–2020	
	N	%	N	%
▶ No	80	32.5	18	14.1
Funding				
▶ Reported	132	53.7	69	54.8
▶ Not reported or not clearly stated	114	46.3	57	45.2
Total	246	100	126	100

%, percentage; AMSTAR, assessment of multiple systematic reviews; n, frequency; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROSPERO, Prospective Register of Systematic Reviews; QUADAS, Quality Assessment of Diagnostic Accuracy Studies; QUOROM, quality of reporting of meta-analysis; STARD, Standards for the Reporting of Diagnostic Accuracy Studies.

procedures. Most usual combinations were the use of two tools or one tool and literature recommendations.

Tools to assess measurement properties of HRQoL instruments

The first 12 columns of [table 2](#) present the characteristics for the identified tools used to assess measurement properties using the last update we are aware of. Tools are reported in order of frequency of use, as pointed out in the last row of the table: (1) ‘COSMIN’, COSMIN initiative^{15 16}; (2) ‘Quality Criteria for Measurement Properties’, Terwee *et al*¹⁷; (3) ‘Attributes and Criteria to assess Health Status and Quality of Life Instruments’, Scientific Advisory Committee Medical Outcomes Trust (SACMOT)^{18 19}; (4) ‘Health Status Measures in Economic Evaluation’, Brazier *et al*^{20 21}; (5) ‘Guidance for Industry PROMs’, Food and Drug Administration (FDA)^{22 23}; (6) ‘Evaluating Patient-based Outcomes Measures for use in clinical trials’, Fitzpatrick *et al*²⁴ (also known as Fitzpatrick’s criteria); (7) ‘International Classification of Functioning’ and ‘International Classification of Functioning for Children and Youth’, WHO²⁵; (8) ‘Evaluating Measures of Patient-Reported Outcomes (EMPRO)’, Spanish Cooperative Investigation Network for Health and Health Service Outcomes Research²⁶; (9) ‘Spinal Cord Injury Criteria’, Spinal Cord Injury Rehabilitation Evidence^{27 28}; (10) ‘Criteria for Assessing the Tools of Disability Outcomes Research’, Andresen²⁹ (also known as Andresen’s tool); (11) ‘CanChild Outcomes Measures’, CanChild Center for Childhood Disability Research³⁰ and (12) ‘Outcomes Measures in Rheumatology Clinical Trials (OMERACT)’, OMERACT initiative.³¹ [Table 2](#) also includes a final column showing the characteristics of Testing Standards by American Educational Research Association, American Psychological Association and National Council on Measurement in Education^{32 33} (hereinafter ‘Testing Standards’) initially published in 1954 and regularly updated every decade using consensus based procedures. The Testing Standards are the source of most of the technical vocabulary for measurement

Table 2 Tools to assess measurement properties, characteristics and comparison to testing standards

Tools	Cosmin	Terwee's criteria	Attributes and criteria	Economic evaluation	Guidance for industry	Fitzpatrick's criteria	ICF ICFCY	EMPRO	SCI criteria	Andresen's tool	Canchild outcomes	Omeract	Testing standards
Development	Delphi	Author criteria	Expert panel	Literature	Consensus	Literature	Expert panel literature	Expert panel literature	Expert panel literature	Literature	Expert panel	Expert panel Delphi	Consensus
Sponsor/s	COSMIN initiative	Author	SACMOT working group	Standing group of health technology	FDA staff	Standing group of health technology	WHO member states	IRYSS committee	SCIRE working group	Author	CanChild centre staff	OMERACT initiative	AERA, APA, NCME
Approval updates	2010, 2018	2007	1996, 2002, 2013	1999, 2017	2006, 2009	1998	2001, 2019*	2008	2008, 2016	2000	1987†, 2004	1992, 1998, 2007, 2014, 2019	1954, 1966, 1974, 1985, 1999, 2014
Items (scoring)	5–18 items/box (+/-?)	8–9 items total (+/-?)	Not item structured (no scoring)	Not item structured (no scoring)	Not item structured (no scoring)	Not item structured (no scoring)	Not item structured (no scoring)	39 items (strongly agree, agree, strongly disagree)	3–5 items/box (++++/+++//++/+/)	Eleven items total (A, B, C)	2–6 items/box (excellent, adequate, poor)	2–5 items/box (Green, amber, red, white)	Not item structured (no scoring)
Measurement properties	Content construct (Int. structure cross-cultural hypotheses test) Criterion (Gold standard)	Content construct (Hypotheses test) Criterion (Gold standard) Floor/Ceiling	Conceptual model	Descriptive (Content Face Construct Preference-based valuation Empirical Criterion)	Conceptual model Content construct (Hypothesis test, discriminant, convergent, known groups)	Use Content/face construct (convergent, discriminant, int. structure) Criterion (Predictive) Cut-score precision	Content (no scoring)	Measurement model Content construct (Hypotheses test) Criterion	Content criterion (concurrent predictive 'discriminant' Clinical utility (consequential validity) Floor/Ceiling)	Conceptual and measurement model Instrument bias Int. structure convergent discriminant	Use scale construction Content construct (Hypotheses test) Criterion (Gold standard)	Content, face construct (Convergent, divergent Criterion (Accuracy) Discrimination (Sensitivity over time and over treatment)	Content response process Int. structure (Dimensions, DIF) Relations to other variables (Hypotheses test, Convergent, Discriminant, criterion, responsiveness Consequences
Validity	Content construct (Int. structure cross-cultural hypotheses test) Criterion (Gold standard)	Content construct (Hypotheses test) Criterion (Gold standard) Floor/Ceiling	Conceptual model Content construct (Hypothesis test, discriminant, convergent, known groups)	Descriptive (Content Face Construct Preference-based valuation Empirical Criterion)	Conceptual model Content construct (Hypothesis test, discriminant, convergent, known groups)	Use Content/face construct (convergent, discriminant, int. structure) Criterion (Predictive) Cut-score precision	Content (no scoring)	Measurement model Content construct (Hypotheses test) Criterion	Content criterion (concurrent predictive 'discriminant' Clinical utility (consequential validity) Floor/Ceiling)	Conceptual and measurement model Instrument bias Int. structure convergent discriminant	Use scale construction Content construct (Hypotheses test) Criterion (Gold standard)	Content, face construct (Convergent, divergent Criterion (Accuracy) Discrimination (Sensitivity over time and over treatment)	Content response process Int. structure (Dimensions, DIF) Relations to other variables (Hypotheses test, Convergent, Discriminant, criterion, responsiveness Consequences
Reliability	Int. consistency measurement error (test retest, agreement) relative measurement error)	Int. consistency reproducibility (Agreement, relative measurement error)	Test retest Inter-rater consistency	Test retest Inter-rater consistency	Test retest Inter-rater consistency	Int. consistency reproducibility (Test retest)	Int. consistency reproducibility (Test retest, inter-rater)	Int. consistency reproducibility (Test retest, inter-rater)	Int. consistency test retest	Int. consistency test retest	Int. consistency intra/inter-rater test retest	Reproducibility test retest	Int. consistency test retest alternate forms scores and decision consistency/accuracy
Fairness													Equivalence of accommodations
Other characteristics									Norms	Norms, standard values	Norms standardisation		Scales, norms, Score comparability
	Interpretability	Interpretability	Interpretability	Interpretability	Interpretability	Interpretability	Interpretability	Interpretability	Interpretability	Burden	Burden	Burden	Test development and revision
	Feasibility	Cultural adaptations	Administration accessible forms	Practicality	Administration accessible forms	Administration accessible forms	Feasibility cultural adaptations	Cultural adaptations	Applicability cultural adaptations	Cultural adaptations	Clinical utility (Feasibility)	Feasibility	
Frequency of use (%)	61 (30.4)	45 (22.4)	33 (16.4)	17 (8.4)	14 (6.9)	14 (6.9)	7 (3.4)	4 (2.0)	2 (1.0)	2 (1.0)	1 (0.5)	1 (0.5)	0

*Updated version at website.

†Reference at 2004.

AERA, American educational research association; APA, American Psychological Association; COSMIN, Consensus-based Standards for the selection of Health Measurement Instruments; DIF, differential item functioning; EMPRO, Evaluating Measures of Patient Reported Outcomes; FDA, Food and Drug Administration; ICF, international classification of functioning; ICFCY, international classification of functioning for children and youth; IRYSS, Investigation Network for Health and Health Service Outcomes Research; NCME, National Council on Measurement in Education; OMERACT, Outcomes Measures in Rheumatology Clinical Trials; SACMOT, Scientific Advisory Committee Medical Outcomes Trust; SCI, spinal cord injury; SCIRE, Spinal Cord Injury Rehabilitation Evidence.

properties in HRQoL instruments, therefore, they will be used as a reference to compare the twelve identified tools. In fact, these standards have already been recommended to establish a unified approach to validity and reliability of results derived from psychometric instruments in clinical medicine, research and education.³⁴

Different methodologies were used to develop the tools. The expert panel consensus and the literature review were the most usual methods, led by steering committees or staff/working groups. The format and structure of these tools also vary. Whereas seven of them were itemised to allow the assignment of quality scores, the other six took the form of standards or guidelines. Tools with an itemised structure were the COSMIN, Quality Criteria for Measurement Properties, EMPRO, SCI Criteria, Criteria for Assessing the Tools of Disability Outcomes Research (Andresen's tool), CanChild Outcomes Measures and OMERACT.

Among all measurement properties considered in Testing Standards, 11 out of the 12 tools recommended to assess the conceptual and measurement model; content, structural, convergent, discriminant, concurrent and predictive validity; responsiveness or sensitivity to change; and internal consistency, test–retest and inter-rater reliability. However, the approach to analyse these measurement properties varied, with examples found in construct validity, criterion validity and reliability. Depending on the tool, the validity of the construct can be evaluated either by hypothesis confirmation in general (eg, COSMIN or EMPRO), or by specific hypothesis based on correlations with other measures, that is, convergent and discriminant validity (eg, Andresen's tool). Criterion validity can be assessed either exclusively by calculating the correlation coefficient with a gold standard (eg, CanChild Outcomes Measures) or by obtaining variously correlation, specificity and sensitivity or predictive values (eg, FDA). Reliability can be analysed either by test retest reliability, inter-rater reliability and internal consistency (eg, FDA), or only by test retest and inter-rater agreement (eg, Economic evaluation). Despite the Testing Standards recommendations, just one tool includes additional criteria to assess consequential validity (SCI), and four assess fairness (eg, accessible forms for subjects with vision impairment or for specific populations) (SACMOT, FDA, SCI and Andresen's tool). None of them includes criteria to assess the validity of response processes. Other HRQoL instrument characteristics, such as feasibility (eg, cost of obtaining a sample), acceptability (eg, suitability from the patient perspective) or burden (eg, the time or effort placed on the administration of the instrument) are assessed instead. Finally, notice that some concepts have changed their place over time. The clearest case is evidence regarding cross-cultural equivalence, which was treated as an additional characteristic of the instruments in most tools released before 2014 (eg, EMPRO or SCI), but was considered a proper measurement property in the COSMIN's 2018 update. It is also considered a measurement property

in Testing Standards where it is included as a particular case of differential item functioning when assessing the internal structure of the instruments (see online supplementary file S3 for more details).

Intended uses of instruments and their association to measurement properties

Some of the differences between tools can be attributed to the fact that they are devoted to the evaluation of instruments developed with different intended uses. For instance, COSMIN aims at assessing the quality of instruments for an evaluative purpose whereas the Economic Evaluation tool aims at the assessment of instruments for analytical purposes. Nevertheless, the relation between the intended use of the instruments and the measurement properties assessed is not usually included in the conclusions of the systematic reviews. Table 3 shows the intended use of instruments, based on the framework proposed by McDowell *et al*.³⁵ and the association to measurement properties that reviewers established in their conclusions. The instruments were most frequently used for evaluation (178, 72.3%) and for assessment of impact of disease on HRQoL (138, 55.1%), either alone or in conjunction. Other purposes were analytic (35, 14.2%), diagnostic (16, 6.5%), descriptive (4, 1.6%) and predictive (2, 0.8%). A total of 6 (2.4%) systematic reviews did not report or did not clearly state the intended use of the instruments. As far as the assessment and conclusions is concerned, only 68 (27.6%) systematic reviews linked the intended use of the instrument to measurement properties. The most common use was evaluative, generally associated to responsiveness, content validity or reliability, for example. When the purpose was the assessment of the impact of disease on HRQoL, the conceptual and measurement model and content validity were usually reported. The analytical purpose involved reporting preference-based valuation (eg, utility scores) and evidence of agreement, and the diagnostic use was linked to known groups validity and test–retest reliability. To better understand these results, some examples are given. First, the evaluative purpose was associated to responsiveness, we found conclusions such as: 'For use in longitudinal studies or clinical practice, where responsiveness is an issue, the Minnesota Living with Heart Failure Questionnaire and the Chronic Heart Failure Questionnaire would be adequate'.³⁶ Second, the intended use was the assessment of the impact of disease on HRQoL, the usual association was to the measurement model and conclusions resembled this one: 'None of the RLS specific QOL measures appears to have been informed by a conceptual model or a conceptual framework. Consequently, none can be considered comprehensive in terms of assessing the full impact of Rest Legs Syndrome on QOL'.³⁷ Third, an example illustrating general conclusions, that is, conclusions that did not associate the intended use of the instrument to any specific measurement properties, was as follows: 'None of the available instruments fulfils the

Table 3 Intended use of instruments and their association to measurement properties

Intended use of instruments identified across the systematic reviews	Frequency	% (over 246)
Evaluative (Change scores pre and poststudies. Effectiveness of an intervention)	178	72.3
Impact of disease on HRQoL (disease symptoms, burden...)	138	55.1
Analytic (health policies. Cost-effectiveness. Funding)	35	14.2
Diagnostic (Distinguish between groups, levels of severity...)	16	6.5
Descriptive (Health measures in surveys. Needs of groups of people)	4	1.6
Predictive (Anticipation of future health status. Risk factors. Risk profiles)	2	0.8
Intended use is no reported or no clearly stated	6	2.4
Conclusions according to the intended use of instruments	n	% (over 246)
Yes, reviewers made specific conclusions	68	27.6
No, reviewers made general conclusions	178	72.4
Measurement properties associated to the intended use of the instrument	n	% (Over 68)
Evaluative		
Responsiveness/Conceptual and Measurement Model/Content validity/Reliability (internal consistency, test retest)/Respondent Burden/Convergent validity/Cross cultural validity	41	60.3
Impact		
Conceptual and Measurement Model/Content validity	29	42.6
Analytic		
Preference-based valuation/agreement	11	16.2
Diagnostic		
Known groups validity/test-retest	7	10.3
Predictive		
Sensitivity and specificity	1	1.5

(%), percentage.
HRQoL, health-related quality of life.

psychometric demands of reliability, validity and responsiveness to serve as a primary outcome measure in clinical trials'.³⁸

DISCUSSION

The present meta-review identified 246 systematic reviews assessing measurement properties of HRQoL instruments in order to analyse the quality of the review

process, describe the most used tools to assess measurement properties and examine how reviewers included the assessment of the quality of HRQoL in their conclusions.

Reporting and methodological quality of the studies

Findings showed how the reporting and methodological quality of systematic reviews has increased over time. Most reviewers reported the search strategy, stated the inclusion and exclusion criteria taking the judgement of two or more independent reviewers into account and used a flow chart to report search outcomes. However, some crucial methodological shortcomings were found. Practices such as registration of the protocol, reporting the detailed search syntax for one database at least, adherence to reporting guidelines, and assessing the reporting and the methodological quality of primary studies were quite sparse even in recent years. As Pussegoda *et al*⁴ suggested, this fact may be related to the perceived time-consuming task of using guidelines or to the lack of information about the most appropriate tool. According to our data, there is still large room for improvement in the assessment of the methodological quality of included studies in order to attend to Terwee *et al*'s warning² of avoiding the risk of presenting biased results, leading to underestimation or overestimation of the quality of an instrument.

Assessment of measurement properties of HRQoL instruments

Assessment procedures of measurement properties of HRQoL instruments were diverse. Most of the reviewers used at least one tool. Nevertheless, there were reviewers that applied their own criteria, followed literature recommendations or applied different ad hoc devised checklists. The use of such diverse procedures is noticeable, even in recent years, when well-accepted tools to assess measurement properties are available.

Our meta-review identified up to twelve tools. Seven of them had an itemised structure, offering a comparable approach to rate the evidence on measurement properties. Length and scoring differed, but also the instrument assessment criteria. Actually, depending on the tool used, the approach to assess properties varied greatly, with potentially serious consequences. The fact that a single measurement property is or isn't required can change the status of quality of the evidence supporting the same measurement instrument. The variety of forms found were in concordance to results from related research, which also highlighted the complexity with regard to definitions of measurement properties.⁶ This complexity is also reflected in the search filter developed by the COSMIN initiative.¹³ They recommend using three filters that sum up more than 100 search terms in order to get sensible and specific results. In addition, and also depending on the tool used, other characteristics, such as feasibility, acceptability and burden were assessed. In spite of the diversity, a shared conclusion can be stated as follows: because these instruments are to be used in the daily practice, their usability should be always balanced with other characteristics considered as proper measurement

properties.^{39 40} For instance, an instrument needs to be long enough to ensure reliability and construct validity, but short enough to ensure the adequate response rate and sample size. Otherwise the instrument intended use and sustainability will be at hazard.³⁹

The differences between tools and their potentially serious consequences on the assessment of the quality of the primary studies may be better addressed in the light of three considerations: the date of publication, the main scientific tradition involved when developing the tools, and the intended uses of the instruments under assessment. Some differences can be simply explained by the date of publication of the tools. As an example, where older tools require specific forms of validity evidence related to external variables such as convergent and discriminant validity, recent tools incorporate the more general view of hypothesis testing. That is, when developing a new use for an instrument, hypotheses should be made regarding the expected relations with other relevant variables in their nomological network and these hypotheses and no other should be tested.³² Regarding the scientific traditions, the assessment of outcomes is a constitutive part of the disciplines of Education and Psychology where the Testing Standards come from. In these contexts, participation is taken for granted as assessment practices result in high stakes decisions such as, for instance, certification or personnel selection. The main concern regarding integrity of the instrument purpose is its fakeability, which could distort the decision-making process, and this would explain the interest in response processes in this field.^{41 42} By contrast, the main objective in the discipline of Medicine is to provide healthcare services. Evaluation of subjective views of patients was a late addition related to the inclusion of HRQoL in the accounting of healthcare outcomes, despite the instruments assessing the patient experience should be acceptable to both patients and clinicians, as Beattie *et al* highlighted.³⁹ Specifically, in the context of disability research, the administrative and respondent burden requires additional consideration. The administrative burden may include the need for a Sign Language interpreter, and the respondent burden includes the length of the questionnaire, which is especially relevant when using HRQoL instruments with cognitively impaired subjects.²⁹ Balancing the traditional psychometric criteria, the practicalities of the instruments and patient preferences is a generic recommendation for health research, but becomes a special obligation for research with people with specific needs.²⁹ Moreover, devising test accommodations or accessible forms when needed is expected to become a required psychometric criterion in the near future, given that it has already been included under the title 'fairness in testing' as a new section next to validity and reliability in the chapter of measurement foundations in the most recent update of Testing Standards.³²

Another criterion is that of economic evaluation, traditionally embedded in providing quantitative judgements able to be integrated into mathematical models such as

those used in calculating quality-adjusted life years and using preference-based methods to obtain their data. Due to that, some very popular measurement properties such as internal structure based on factor analysis are not relevant and thus not considered in their tools. In this tradition, the main concern regarding the integrity of the instrument purpose is whose values should be considered when determining preferences and how well the preferences of patients and decision makers are likely to conform to the main assumptions of the utility models.^{20 21}

Intended uses of instruments and their association to measurement properties

In our view, considering in the first place the intended use of the HRQoL instrument would help to reconcile the different requirements included in each tool. Tools for evaluating the measurement quality of instruments should be adapted or extended according to the different intended uses of these instruments, such as evaluative, impact of disease, analytic, diagnostic, descriptive or predictive. Notice that depending on the intended use of the measure, some domains of validity and reliability may be of greater or lesser relevance.^{6 16} For instance, an instrument developed to assess longitudinal changes should demonstrate high responsiveness,⁶ but if used for diagnostic purposes, it should be able to distinguish among individuals or groups,⁶ that is, known groups validity. Another example is the internal consistency reliability based on interitem relationships that may be not relevant for a preference-based instrument but is relevant for an instrument based on a unidimensional measurement model. However, our data showed that only a few authors established a clear link in their recommendations between the intended use of the measure and the reported evidence of measurement properties. The vast field of HRQoL offered a plethora of instruments but, as most reviewers did not take the intended use of the instrument into account, the overall rating of measurement properties was not consistent and thus the instrument may or may not have been adequate for its intended use. Because the evaluation and improvement of quality of life is considered a public health priority,¹⁴ we strongly encourage researchers to assess the quality of measurement properties of HRQoL instruments according to the intended use of the measure. Otherwise, there is a serious risk of biased results, which could lead to underrating the quality and suitability of the instrument.

CONCLUSIONS

The quality of the systematic review process has been increasing over time, but it should still improve with regard to the prospective registration of protocol, and with respect to the adoption of guidelines to improve both the methodological and reporting quality of the reviews. In the specific context of systematic reviews of measurement instruments, enhancing the quality of the

process also involves the assessment of measurement properties by using a standardised tool. The selection of the most suitable tool may be addressed according to the coverage of the appraised measurement properties, but also according to other important criteria, such as the intended use of the HRQoL instruments, the format of the tool and whether it assesses both usability (eg, feasibility or burden) and accommodation (or accessible forms). First, the assessment methodology should be adapted when necessary, establishing the relation between the intended use of the HRQoL instruments and the measurement properties assessed. Second, to standardise the review process, the tool's format should be itemised offering a comparable approach to rate the evidence on measurement properties. Those tools that take the form of guidelines, such as the SACMOT or the economic evaluation would be considerably upgraded if the structure is reconverted, since the current format only allows description rather than critical appraisal of the quality of an instrument, and furthermore, it complicates comparison of results. Lastly, because systematic reviews on measurement properties aim to help professionals to select the best instrument for a clinical scenario, the feasibility, patient's preferences, administrator and respondent burden, and the accommodations (or accessible forms) should be addressed and evaluated. Otherwise the suitability and the intended use of instruments might be compromised, especially in the context of disability research. Tools identified in our meta-review that meet most of these criteria are the COSMIN, EMPRO, SCI criteria, Andresen's tool, CanChild Outcomes and OMERACT, since all of them cover a wide range of measurement properties, offer an item structure, and assess the usability of instruments.

Special mention is due to the COSMIN, the most widespread and comprehensive tool to assess measurement properties of health instruments designed for an evaluative purpose. The COSMIN standards were developed in a Delphi study⁴³ aiming to improve the selection of the most appropriate health instrument for a clinical scenario. The most recent version of the COSMIN consists of a manual for conducting systematic reviews of health instruments, providing different steps with respect to the literature search process, the assessment of measurement properties and feasibility of instruments, and the evaluation of the risk of bias (RoB) of studies according to the Cochrane methodology.¹⁶ Additionally, the COSMIN initiative recently developed a guideline exclusively focused on assessing the content validity of health instruments, considered the most important property to ensure the adequate reflection of the construct measured.^{44 45} In the light of these considerations, we strongly recommend the application of the latest version of the COSMIN to conduct high-quality systematic reviews on measurement properties of health instruments for an evaluative purpose, or for other purposes with appropriate adaptation.

Despite COSMIN's many strengths, our analysis of the other assessment tools and measurement standards allow

us to suggest future lines of work on this tool. First, the current format of COSMIN is fairly complex, requiring high expertise in the field of psychometrics and specific training for its proper application. The reporting of the inter-rater agreement coefficients when reviewers use the last version of COSMIN may provide useful data about its reliability. Second, consideration should be given to the testing standards recommendation on the inclusion of the assessment of fairness (ie, evaluation of accessible forms for specific populations). Third, the feasibility of the measurement instruments, merely described in COSMIN, and their burden, should be properly rated, with examples found in EMPRO or Andresen's tool. Fourth, it must be considered that the RoB evaluation of studies is itself a productive field of research with a long tradition, with specific tools that have been developed for different research questions and study designs. Examples might be found in the Cochrane Collaboration's Tool for Assessing the Risk of Bias of Clinical Trials,⁴⁶ the Newcastle Ottawa Scale⁴⁷ for non-randomised studies, or the Quality Assessment Tool for Cohort Studies.^{48 49} From our point of view, the COSMIN proposal could also be simplified and improved by guiding the reviewers towards the identification of the most appropriate RoB assessment tools instead of developing their own RoB appraisal guidelines, taking advantage of knowledge and innovations in that field of research.

And last, but not least, improving the quality of systematic reviews encompasses researchers, sponsors and promoters, but also journals, which should require full compliance with reporting and methodological guidelines, and the use of assessment tools.

Twitter Jaume Vives @VivesJ_Research and Josep-Maria Losilla @jmlosilla

Contributors All authors meet the criteria recommended by the International Committee of Medical Journal Editors, ICMJE. All authors made substantial contributions to conception and design, piloted the inclusion criteria and provided the direction of the data extraction and analysis. SL drafted the article and JV, CV and J-ML critically revised the draft for important intellectual content. All authors agreed on the final version.

Funding This work was supported by the Grant PGC2018-100675-B-I00, Spanish Ministry of Science, Innovation and Universities (Spain).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Data are available on reasonable request to the authors.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Sonia Lorente <http://orcid.org/0000-0002-5494-3325>

REFERENCES

- 1 Mokkink LB, Terwee CB, Stratford PW, *et al.* Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Qual Life Res* 2009;18:313–33.
- 2 Terwee CB, Prinsen CAC, Ricci Garotti MG, *et al.* The quality of systematic reviews of health-related outcome measurement instruments. *Qual Life Res* 2016;25:767–79.
- 3 Prinsen CAC, Mokkink LB, Bouter LM, *et al.* COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
- 4 Pussegoda K, Turner L, Garritty C, *et al.* Identifying approaches for assessing methodological and reporting quality of systematic reviews: a descriptive study. *Syst Rev* 2017;6:1–12.
- 5 Pussegoda K, Turner L, Garritty C, *et al.* Systematic review adherence to methodological or reporting quality. *Syst Rev* 2017;6:1–14.
- 6 Rosenkoetter U, Tate RL. Assessing features of psychometric assessment instruments: a comparison of the COSMIN checklist with other critical appraisal tools. *Brain Impair* 2017:1–16.
- 7 Reeve BB, Wyrwich KW, Wu AW, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889–905.
- 8 Lorente S, Vives J, Viladrich C, *et al.* Tools to assess the measurement properties of quality of life instruments : a meta-review protocol. *BMJ Open* 2018;8:1–4.
- 9 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264–9.
- 10 Liberati A, Altman DG, Tetzlaff J, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med* 2009;151:W65–94.
- 11 McGowan J, Sampson M, Salzwedel DM, *et al.* PRESS peer review of electronic search strategies: 2015 guideline explanation and elaboration (PRESS E&E). *Cadth Methods Guidel* 2016:40–6.
- 12 McGowan J, Sampson M, Salzwedel DM, *et al.* PRESS peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol* 2016;75:40–6.
- 13 Terwee CB, Jansma EP, Riphagen II, *et al.* Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115–23.
- 14 Secretary's Advisory Committee on National Health Promotion and Disease Prevention Objectives. Health-related quality of life and well-being, 2010. Available: <https://www.healthypeople.gov/sites/default/files/HRQoLWBFFullReport.pdf>
- 15 Mokkink L, Terwee C, Patrick D, *et al.* The COSMIN checklist manual, 2012. Available: <http://www.cosmin.nl/images/upload/files/COSMINchecklistmanualv9.pdf>
- 16 Mokkink LB, Prinsen C, Patrick D, *et al.* COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual* 2018:1–78.
- 17 Terwee CB, Bot SDM, de Boer MR, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
- 18 Aaronson N, Alonso J, Burnam A, *et al.* Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–215.
- 19 Lohr KN, Aaronson NK, Alonso J, *et al.* Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;18:979–92.
- 20 Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999;3:1–164.
- 21 Brazier J, Ratcliffe J. Measurement and Valuation of Health for Economic Evaluation. In: *International encyclopedia of public health*. Elsevier, 2010: Vol. 4. 586–93.
- 22 Department of Health and Human Services. *Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance*. Vol. 20, 2006.
- 23 Department of Health and Human Services. *Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims*, 2009.
- 24 Fitzpatrick R, Davey C, Buxton MJ, *et al.* *Evaluating patient-based outcome measures for use in clinical trials*. Vol. 2, 1998.
- 25 World Health Organization. International classification of functioning (ICF), 2016. Available: www.who.int/classifications/icf/en/ [Accessed 6 Feb 2020].
- 26 Valderas JM, Ferrer M, Mendivil J, *et al.* Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. *Value Health* 2008;11:700–8.
- 27 SCIRE. Spinal cord injury rehabilitation evidence. Available: <https://scireproject.com>
- 28 Johnston MV, Graves DE. Towards guidelines for evaluation of measures: an introduction with application to spinal cord injury. *J Spinal Cord Med* 2008;31:13–26.
- 29 Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil* 2000;81:S15–20.
- 30 Law M. Outcome measures rating form guidelines, 2004. Available: <https://www.canchild.ca/system/tenon/assets/attachments/000/000/371/original/measguid.pdf>
- 31 OMERACT. *Instrument selection for core outcome measurement sets*. OMERACT Handbook, 2019. <https://omeracthandbook.org/handbook>
- 32 APA, AERA, NCME. *Standards for educational and psychological testing*. American Educational Research Association, 2014.
- 33 APA, AERA, NCME. *Standards for educational and psychological testing*. American Educational Research Association, 1999.
- 34 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–166.e16.
- 35 McDowell I, Spasoff RA, Kristjansson B. On the classification of population health measurements. *Am J Public Health* 2004;94:388–93.
- 36 Garin O, Ferrer M, Pont A, *et al.* Disease-specific health-related quality of life questionnaires for heart failure: a systematic review with meta-analyses. *Qual Life Res* 2009;18:71–85.
- 37 Speight J, Howarth A. Quality of life in restless legs syndrome: a systematic review of clinical trials and a critical review of instruments. *Patient* 2010;3:185–203.
- 38 Chassany O, Holtmann G, Malagelada J, *et al.* Systematic review: health-related quality of life (HRQOL) questionnaires in gastro-oesophageal reflux disease. *Aliment Pharmacol Ther* 2008;27:1053–70.
- 39 Beattie M, Lauder W, Atherton I, *et al.* Instruments to measure patient experience of health care quality in hospitals: a systematic review protocol. *Syst Rev* 2014;3:4.
- 40 Lorente S, Losilla J-M, Vives J. Instruments to assess patient comfort during hospitalization: a psychometric review. *J Adv Nurs* 2018;74:1001–15.
- 41 Ferrando PJ, Anguiano-Carrasco C. A structural model-based optimal Person-Fit procedure for identifying Faking. *Educ Psychol Meas* 2013;73:173–90.
- 42 Ferrando PJ, Anguiano-Carrasco C. A structural equation model at the individual and group level for assessing faking-related change. *Struct Equ Model* 2011;18:91–109.
- 43 Mokkink LB, Terwee CB, Patrick 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. *Qual Life Res* 2010;19:539–49.
- 44 Terwee CB, Prinsen CA, Chiarotto A. *COSMIN methodology for assessing the content validity of PROMs: user manual*. Vol. 120, 2018.
- 45 Terwee CB, Prinsen CAC, Chiarotto A, *et al.* COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res* 2018;27:1159–70.
- 46 Higgins JPT, Altman DG, Gøtzsche PC, *et al.* The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928–9.
- 47 Wells G, Shea B, O'Connell D, *et al.* The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses, 2000. Available: www.ohri.ca/programs/clinical_epidemiology/oxford.asp
- 48 Jarde A, Losilla J-M, Vives J, *et al.* Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *Int J Clin Heal Psychol* 2013;13:138–46.
- 49 Jarde A, Losilla J-M, Oliveras I. *Quality assessment tool for cohort studies (Q-COH II) user's manual*, 2014: 1–13.