www.nottingham.ac.uk/praised/index.aspx

The Promoting Activity, Independence and Stability in Early Dementia (PrAISED) research programme is a NIHR funded project that has been designed to help people with mild cognitive impairment or early stage dementia to remain healthier and more independent for longer. We have designed an activity and exercise programme consisting of a combination of exercises, activities of daily living and memory strategies to help improve and maintain individual physical and mental health.

PrAISED Discussion Paper Series

ISSN 2399-3502

Issue 5, July 2019

A systematic review of measures of apathy for older adults: validity, reliability and conceptualisation of apathy.

Burgon C[1,2], Goldberg S[2], van der Wardt V[1] and Harwood RH[2,3]

1. Division of Rehabilitation and Ageing, University of Nottingham, UK
2. School of Health Sciences, University of Nottingham, UK
3. Nottingham University Hospitals, UK

Address for Correspondence:
Clare Burgon, Division of Rehabilitation and ageing, University of Nottingham, B114, B floor, Queen's Medical School, Queen's Medical Centre, Nottingham, NG7 2UH. Email: clare.burgon@nottingham.ac.uk

1. Title: A systematic review of measures of apathy for older adults: validity, reliability and conceptualisation of apathy.

2. Registration

In accordance with the PRISMA guidelines, this systematic review protocol was registered with PROSPERO (ID: CRD42018094390 on 4th June 2018).

3. Authors

Burgon. C.[1,2], Goldberg. S.[2], Van der Wardt. V.[1], Harwood. R.H.[2,3]

1. Division of Rehabilitation and Ageing, University of Nottingham, UK
2. School of Health Sciences, University of Nottingham, UK
3. Nottingham University Hospitals, UK

3.1. Contact Details

Address for correspondence: Clare Burgon, Division of Rehabilitation and ageing, University of Nottingham, B114, B floor, Queen's Medical School, Queen's Medical Centre, Nottingham, NG7 2UH.

Email: clare.burgon@nottingham.ac.uk

3.2. Contributions

CB is the guarantor. CB designed the work proposed in this protocol and wrote the protocol. RH, SG, and VvdW supervised this work and provided feedback on the proposed design and writing.

4. Amendments

If this protocol requires amending, the date of each amendment will be recorded, and change will be described and explained. Changes will be recorded in PROSPERO, but will not be incorporated into this protocol.

5. Support

5.1. Sources of financial and other support

The School of Health Sciences and the Division of Rehabilitation and Ageing at the University of Nottingham have provided financial support and resources to for CB's PhD project, which includes this review.

5.2. Sponsor

This review forms part of CB's PhD study. There is no sponsor.

5.3. Role of sponsor or funder

No sponsor or funder has had an input into the planned works outlined in this protocol.

6. Introduction

6.1. Rationale

Apathy has been defined as a lack of motivation, underpinned by reduced: goal-orientated behaviour, emotional responsiveness, and goal-directed cognition (Marin, 1991). However, there is no definitive consensus on what constitutes apathy: the involvement of motivation has been questioned, with some arguing apathy should be defined as impaired initiation (Stuss, Van Reekum, & Murphy, 2000), and others defining apathy as simply a quantitative reduction in goal-orientated action (Levy & Dubois, 2006). Moreover, it has been suggested that reduced emotional responsiveness is not a necessary part of the apathy construct (Starkstein & Leentjens, 2008).

Despite the lack of consensus, interventions to reduce apathy have been developed and tested (Cipriani, Lucetti, Danti, & Nuti, 2014) and apathy has been identified as a priority area for dementia research (Pickett et al., 2018). However, there is no gold-standard measure of apathy (Clarke et al., 2011). Whilst clinician interview may be the preferred method of identifying apathy, there is no formal diagnostic category for apathy, and quantifiable measurement scales are recommended for research into clinical interventions (Cummings et al., 2015). The assessment of apathy is complicated by the overlap between symptoms of other disorders. Older adults in particular may show diminished goal directed behaviour, due to pain, poor mobility, sensory loss or cognitive impairment (Marin, 1990).

Clarke et al (2011) and Weiser and Garibaldi (2015) have previously reviewed measures of apathy, but these were not systematic reviews. One systematic review of apathy measures has been published (Radakovic, Harley, Abrahams, & Starr, 2015), which examined measures developed for people with neurodegenerative conditions between 1980 and 2013. Measures of apathy developed for other relevant groups, such as people with mild cognitive impairment, were not included in the search strategy, participants' ages were not reported and a broad eligibility criteria of 18 years and above was used, meaning that findings may not be directly applicable to older adults. Radakovic et al. (2015) utilised published criteria for methodological quality, however these criteria did not utilize a scoring system, did not assess the adequacy of the measurement properties themselves, and were developed for studies of diagnostic accuracy. Since then, the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) programme of work has

published guides for conducting and reporting systematic reviews of health measures, which includes quality criteria for good measurement properties and methodological quality standards, to support the systematic and standardized critical examination of outcome measures (C. A. C. Prinsen et al., 2018). COSMIN is compatible with PRISMA guidelines, and provides Grading of Recommendations, Assessment, Development and Evaluation (GRADE) criteria, modified for the assessment of the quality of evidence specific for outcome measures (C. A. C. Prinsen et al., 2018). A detailed guideline for the standardized assessment of content validity has recently been published which can be used alongside the COSMIN guide for conducting systematic reviews (C. B. Terwee et al., 2018). The aforementioned review studies did not consistently report or assess the content validity of apathy measures. Content validity of measurement tools is considered the most important measurement property, as this affects other properties, and ultimately, whether a measure is appropriate (C. B. Terwee et al., 2018). A more detailed analysis of how the different measures conceptualize apathy is particularly important given the lack of consensus about what constitutes apathy.

6.2. Objectives

To identify measures of apathy developed or validated in an older adult population.

To assess and compare the measurement properties and characteristics of the identified apathy measures, including the assessment of their quality using COSMIN criteria for good measurement properties, and the quality of this evidence, using the COSMIN risk of bias checklist, and modified GRADE criteria.

To assess and compare how the identified measures conceptualise apathy.

7. Methods

7.1. Eligibility criteria

Study design

Any study that aims to develop or to assess the measurement properties of a measure of apathy will be included. Studies reporting multi-domain measures or inventories that assess other concepts will only be included if there is an apathy sub-scale or sub-score and results of its reliability or validity are reported separately (i.e. not merged with other concepts). Studies that assess apathy in specific contexts, such as health behaviours (e.g. exercise, health eating), academic achievement, and job performance, will be excluded as we are only interested in measures of generalised apathy. Only published full texts will be included. Abstracts, such as poster abstracts, review articles, commentaries, letters and editorials will be excluded, however, where they report the development or validation of a measure, the corresponding author will be contacted to determine if a full text article is available.

Participants

Studies of participants living in the community, aged 65 or above, regardless of the presence or absence of psychiatric or neurodegenerative disorders, will be included. This is with the exception of studies of apathy measures for people with disorders that are dependent on external circumstances, such as post-traumatic stress disorder, substance use disorders, and post-natal depression. Measures of apathy developed for these populations are likely to developed with these circumstances in mind (e.g. trauma, substance use, having a baby), and may not be generalizable to people outside of the circumstance for which it was designed. Studies that include participants under the age of 65 will be included if they have a median or mean age of 65 or older, or if results relevant to this review are reported separately for this age group. Where proxy-reports are used, the age of the participants (i.e. the person who is the focus of the measure) will be the age that must meet the criteria. At least the majority of participants must live in the community (over 50%), or data relevant to this review must be reported separately for a sub-sample of community-dwellers. If age and residential status of sub-samples (e.g. healthy participants, mild cognitive impairment and dementia) are described separately, but results are reported together, these characteristics will be calculated for the overall sample, and checked against review criteria. However, participants described as ‘controls’ will not be included in this calculation, as this indicates the study is not aiming to assess reliability or validity of the measure for this group. Where insufficient data is reported for age or living setting, and suitable information cannot be obtained from the corresponding author, studies will be included, and their suitability for this population will be further discussed in the review.

Languages

Measures administered in any language will be included. Where possible, articles published in a language other than English will be translated. However, due to resource limitations, it may not be possible to obtain translations of all articles. Any relevant articles listed in languages that are not published in English, and could not be translated, will be provided in an appendix.

Year of Publication

No restrictions will be placed on the year of publication.

7.2. Information Sources

MEDLINE (In-Process, Other Non-Indexed Citations and 1946 onwards, via Ovid) EMBASE (1980 onwards, via Ovid), PsycINFO (1806 onwards, via Ovid) and CINAHL (1937 onwards, CINAHL Plus with Full Text) will be searched using the specified search strategy. The reference lists of the included studies, and of any relevant review articles, will be scanned for relevant publications. The search is due to be completed by April 2018.

7.3. Search Strategy

The search strategy will be developed by the lead author and checked by an expert in the field (RH) and a subject librarian. The COSMIN search strategy for health related outcome measures (Caroline B. Terwee, Jansma, Riphagen, & de Vet, 2009), adapted for MEDLINE (via Ovid) will be used for the first part of the search strategy. No search limits will be applied to study design, date, or language. The search will be limited to adult human participants (aged 18-19 and over). Keyword searches will be applied to titles and abstracts, and medical subject headings will be applied where possible. See 8 for the draft MEDLINE (via Ovid) search strategy. Once the MEDLINE search strategy is finalized, the subject headings and syntax will be adapted to suit the other databases. For EMBASE and CINAHL the corresponding COSMIN translations will be used, and the search strategy for PsycINFO will be adapted from the MEDLINE version by CB.

7.4. Study Records

Data management

Data will be retrieved, stored, de-duplicated, and sorted using Endnote X8. Data screening and data extraction tables will be created and completed in Microsoft Excel 2016. Duplicate publications will be further identified by comparing studies of the same measurement tool by their author name, and sample size where necessary, to prevent double counting and thus introduction of bias to results. However, validation of the same measure in a different context, sample, language, or time, or papers reporting a different measurement property will be included as separate studies.

Selection process

The lead review author will screen the titles and abstracts of articles to assess whether they meet the eligibility criteria, if deemed eligible or uncertain of eligibility, included for further review. Further review of full text articles will be independently assessed against eligibility criteria by CB, and a second reviewer will independently assess a randomly selected 10% of articles, as recommended by Boland et al., for circumstances in which resources are limited (2017). Any articles for which there is disagreement between reviewers will be discussed. If questions of eligibility remain, the articles will be assessed by a third reviewer.

The number of excluded articles will be recorded at each stage, and reasons for exclusion of full text articles will be stated. These will be presented in a PRSIMA flow diagram. Reviewers will not be blinded to journals or author information due to resource limitations.

Data collection process

Data extraction will be conducted electronically by CB and will be verified by a second reviewer. Data will be extracted into a data extraction table in Microsoft Excel 2016. Where missing information is encountered, corresponding authors will be contacted for this information. See Appendix C for the proposed data extraction form.

7.5. Data Items

For each study included in the review, data relating to study characteristics, participant characteristics, and measurement properties will be extracted. Study characteristics includes the design, setting, number of participants, comparator measure, percentage and handling of missing items. Participant characteristics will include the age, sex, ethnicity, cognitive ability, and disease status, duration and severity. Measurement properties of reliability (including internal consistency and test-retest reliability), validity (including structural validity, clinical relevance, and content validity) will be recorded, as well as floor and ceiling effects. Measurement characteristics will be considered for each included measure. These will be ascertained from the included studies, but also from the original development article (where not already included) and other relevant sources such as administration manuals, and correspondence with study authors. These characteristics refer to the number of, domains of and a list of the individual items included, the conceptual model and definition of apathy applied, administration time, licensing and cost information, mode of administration, recall period, response options and scoring system. Due to resource limitations, information that is only accessible through the purchasing of a licence not already held by the study team will not be obtained.

7.6. Outcome and Prioritization

COSMIN guidelines for the recommendation of measurement tools (C. A. C. Prinsen et al., 2018) will be followed. This criteria enables the evaluation of the suitability of each measure based on the overall quality of its measurement properties and quality of evidence. In line with COSMIN guidelines, the original development study of all included measures will be assessed to ascertain the quality of measure development, and included studies will be assessed for content validity in relation to the target population of this review, rather than the population of the published article.

7.7. Risk of bias in individual studies

Risk of bias in individual studies will be examined using the COSMIN risk of bias checklist (Mokkink et al., 2017), which assesses the risk of bias, for each measurement property, for each a study. This information will be presented in a data summary table. See Appendix D for further details regarding the risk of bias checklist.

7.8. Data synthesis

All studies meeting the eligibility criteria will be summarised using a narrative synthesis, and presented in a summary of findings table. For each measure, the measurement properties reported in the corresponding studies will be summarised, and, where possible, the quality of this overall data will be assessed using COSMIN quality criteria (see Appendix E). COSMIN procedure for the recommendations of

measures in systematic reviews (C. A. C. Prinsen et al., 2018) will be used to guide any recommendations made.

7.9. Meta-bias(es)

It is not recommended to assess publication bias in systematic reviews regarding measurement tools and their properties, as there is no common system or database in place for registering these studies against which to check (C. A. C. Prinsen et al., 2018).

7.10. Confidence in cumulative evidence

A modified GRADE approach (C. A. C. Prinsen et al., 2018) will be applied to assess the quality of the cumulative evidence for each property of each measure (see Appendix F). This criteria differs slightly for content validity (see: C. B. Terwee et al., 2018).

8. References

Boland, A., Cherry, M. G., & Dickson, R. (2017). *Doing a systematic review : a student's guide* (2nd ed.).

Cipriani, G., Lucetti, C., Danti, S., & Nuti, A. (2014). Apathy and dementia. Nosology, assessment and management. *J Nerv Ment Dis, 202*(10), 718-724. doi:10.1097/nmd.0000000000000190

Clarke, D. E., Ko, J. Y., Kuhl, E. A., van Reekum, R., Salvador, R., & Marin, R. S. (2011). Are the available apathy measures reliable and valid? A review of the psychometric evidence. *Journal of psychosomatic research, 70*(1), 73-97.

Cummings, J., Friedman, J. H., Garibaldi, G., Jones, M., Macfadden, W., Marsh, L., & Robert, P. H. (2015). Apathy in Neurodegenerative Diseases:Recommendations on the Design of Clinical Trials. *Journal of Geriatric Psychiatry and Neurology, 28*(3), 159-173. doi:10.1177/0891988715573534

Levy, R., & Dubois, B. (2006). Apathy and the Functional Anatomy of the Prefrontal Cortex–Basal Ganglia Circuits. *Cerebral Cortex, 16*(7), 916-928. doi:10.1093/cercor/bhj043

Marin, R. S. (1990). Differential diagnosis and classification of apathy. *Am J Psychiatry, 147*(1), 22-30. doi:10.1176/ajp.147.1.22

Marin, R. S. (1991). Apathy: a neuropsychiatric syndrome. *The Journal of neuropsychiatry and clinical neurosciences*.

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2017). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res*. doi:10.1007/s11136-017-1765-4

Pickett, J., Bird, C., Ballard, C., Banerjee, S., Brayne, C., Cowan, K., . . . Walton, C. (2018). A roadmap to advance dementia research in prevention, diagnosis, intervention, and care by 2025. *Int J Geriatr Psychiatry*. doi:10.1002/gps.4868

Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*. doi:10.1007/s11136-018-1798-3

Prinsen, C. A. C., Vohra, S., Rose, M. R., Boers, M., Tugwell, P., Clarke, M., . . . Terwee, C. B. (2016). How to select outcome measurement instruments for outcomes included in a “Core Outcome Set” – a practical guideline. *Trials, 17*(1), 449. doi:10.1186/s13063-016-1555-2

Radakovic, R., Harley, C., Abrahams, S., & Starr, J. M. (2015). A systematic review of the validity and reliability of apathy scales in neurodegenerative conditions. *27*(6), 903-923. doi:10.1017/S1041610214002221

Starkstein, S. E., & Leentjens, A. F. (2008). The nosological position of apathy in clinical practice. *J Neurol Neurosurg Psychiatry, 79*(10), 1088-1092. doi:10.1136/jnnp.2007.136895

Stuss, D. T., Van Reekum, R., & Murphy, K. J. (2000). Differentiation of states and causes of apathy *The neuropsychology of emotion.* (pp. 340-363). New York, NY, US: Oxford University Press.

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. W. (2009). Development of a methodological PubMed search filter for finding studies on

measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115-1123. doi:10.1007/s11136-009-9528-5

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., . . . Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. doi:10.1007/s11136-018-1829-0

Weiser, M., & Garibaldi, G. (2015). Quantifying motivational deficits and apathy: A review of the literature. *European Neuropsychopharmacology, 25*(8), 1060-1081. doi:https://doi.org/10.1016/j.euroneuro.2014.08.018

This paper presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Reference Number RP-PG-0614-20007). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Appendix A. MEDLINE search strategy

	Search terms
1	(instrumentation or methods).sh.
2	(Validation Studies or Comparative Study).pt.
3	exp Psychometrics/
4	psychometr*.ti,ab.
5	(clinimetr* or clinometr*).tw.
6	exp "Outcome Assessment (Health Care)"/
7	outcome assessment.ti,ab.
8	outcome measure*.tw.
9	exp Observer Variation/
10	observer variation.ti,ab.
11	exp Health Status Indicators/
12	exp "Reproducibility of Results"/
13	reproducib*.ti,ab.
14	exp Discriminant Analysis/
15	(reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or "internal consistency").ti,ab.
16	(cronbach* and (alpha or alphas)).ti,ab.
17	(item and (correlation* or selection* or reduction*)).ti,ab.
18	(agreement or precision or imprecision or "precise values" or test-retest).ti,ab.
19	(test and retest).ti,ab.
20	(reliab* and (test or retest)).ti,ab.
21	(stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intratester or intra-tester or interobserver or inter-observer or intraobserver or intra-observer or intertechnician or inter-technician or intratechnician or intra-technician or interexaminer or inter-examiner or intraexaminer or intra-examiner or interassay or inter-assay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-individual or interparticipant or inter-participant or intraparticipant or intra-participant or kappa or kappa's or kappas or repeatab*).ti,ab.
22	((replicab* or repeated) and (measure or measures or findings or result or results or test or tests)).ti,ab.
23	(generaliza* or generalisa* or concordance).ti,ab.
24	(intraclass and correlation*).ti,ab.
25	(discriminative or "known group" or factor analysis or factor analyses or dimension* or subscale*).ti,ab.
26	(multitrait and scaling and (analysis or analyses)).ti,ab.
27	(item discriminant or interscale correlation* or error or errors or "individual variability").ti,ab.
28	(variability and (analysis or values)).ti,ab.
29	(uncertainty and (measurement or measuring)).ti,ab.
30	("standard error of measurement" or sensitiv* or responsive*).ti,ab.
31	((minimal or minimally or clinical or clinically) and (important or significant or detectable) and (change or difference)).ti,ab.
32	(small* and (real or detectable) and (change or difference)).ti,ab.

33	(meaningful change or "ceiling effect" or "floor effect" or "Item response model" or IRT or Rasch or "Differential item functioning" or DIF or "computer adaptive testing" or "item bank" or "cross-cultural equivalence").ti,ab.
34	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11 or 12 or 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33
35	exp APATHY/
36	apath*.mp
37	amotivat*.ti,ab.
38	diminished motivation.ti,ab.
39	diminished interest.ti,ab.
40	lack of interest.ti,ab.
41	diminished initiat*.ti,ab.
42	lack of initiat*.ti,ab.
43	lack of motivation.ti,ab.
44	emotional* blunt*.ti,ab.
45	abulia.ti,ab.
46	anhedonia.ti,ab.
47	exp Anhedonia /
48	frontal symptom*.ti,ab.
49	emotional responsiv*.ti,ab.
50	asocial*.ti,ab.
51	avolition*.ti,ab.
52	lassitude.ti,ab.
53	35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 52
54	34 and 53
55	limit 54 to "all adult (19 plus years)"

Appendix B. Screening and selection tool

Identifier								Type of article		Aim / Design	
								Include if	Exclude if	Include if	
								full text original research	publication is an abstract (e.g. poster) / review article / commentary / letter / editorial etc.	Aim to assess at least one measurement property (reliability or validity) of a measure	Aim is to develop or adapt (e.g. change in purpose, length, or population) a measure
Stage identified (original search or reference lists)											
Number (for randomisation)											
Name of measure											
Title of paper											
Year											
Author											
DOI											

	Participants age			Construct				Exclude if		
	Exclude if	Include if			Exclude if	Include if				
	no adults aged 65 or over	adult participants where results are reported separately for those with a median/mean age of 65 &over	adult participants, with a median/ mean age of 65 or more[1]	all participants are adults aged 65 or more	measure of another construct (or measures apathy as part of another construct without reporting on validity or reliability of an apathy subscale)	Apathy is part of a scale which considers apathy as a distinct concept, and results are reported separately for the apathy subscale in at least one relevant analysis	Independent measure of apathy	Outcome study (e.g. RCT)	Does not aim to develop, validate or assess a measure	

	Residential status				Overall decision			Notes
		Include if		Exclude if				
	participants mean age 64 or under & results not reported separately	At least the majority of participants live in the community - includes retirement villages, and independent (but not assisted) living communities[1]	Results are reported separately for participants living in the community	All participants are not living in the community	Less than half of participants are living in the community and results are not reported separately	**Undecided** (awaiting translation[2])	**Incl**	**Excl**	**Comments & Notes**

[1] Where age / residential status is reported separately for different sub-samples, but results are reported together, consider the overall mean /majority of all sub-samples, except sub-samples described as controls.

[2]For studies published in a language other than English, they will be listed in a separate table, and translations will be sought for these studies. Where this is not possible, untranslated studies will be provided in an appendix in the systematic review.

Appendix C. Data Extraction Tables

Table 1. Data extraction, to be completed for each study, and for each sub-analysis where appropriate.

Identifier						Interpretability				
Name of measure	Title of study	Year	Author	DOI	Number (for randomisation)	% of missing items	% of missing total scores	Floor/ ceiling effects: % min score; % max score	Apathy scores & change scores for relevant (sub)groups (e.g. cognitive impairment vs healthy controls) *Mean (SD), range or similar*	minimal important change or minimal important difference

	Study characteristics								
	Country from which research was conducted	Further description of measure if needed (e.g. changes to original)	Measurement properties assessed (i.e. relevant COSMIN boxes to complete)	N (in each sub-analysis where appropriate)	eligibility criteria	Target population	Setting (location, time, context)	Sampling method	Design	Information on response shift

	Participant Descriptives						Content validity	Psychometric properties
Language of measure	**Age** *Mean (SD), Range or similar*	**Ethnicity**	**Distribution of sex** *% M/ F*	**Disease characteristics** (disease status, severity, duration)	**Cognitive status** *Mean (SD), range of MMSE or similar*	**Residential status** *Type and distribution e.g. % at home in community*	**Content validity** *Describe method briefly, e.g. asking stakeholders (e.g. patients, carers, experts) about the relevance, comprehensiveness, comprehensibility of the measure*	**Structural validity -** (e.g. Comparative Fit Index or Tucker Lewis Index)

	Internal consistency (e.g. chronbach's alpha)	**Reliability** (e.g. intra-class correlation co-efficient or weighted Kappa)	**Measurement error** (Standard error of measurement, Smallest Detectable change, Limits of Agreement, or % agreement)	**Hypothesis testing / Construct validity : comparisons** Describe all comparator instruments used	**Hypothesis testing / Construct validity** (convergent and discriminative; cross-sectional data only) Describe statistical method, result for each relevant comparator measure)	**Cross cultural validity/ measurement invariance.** Differences between group factors	**Responsiveness** (Longitudinal data only. Compared to another measure, compared across groups, or before & after intervention)

For the content validity and psychometric property sections of this table, where the study did not attempt to assess these properties, 'n/a' will be recorded.

Note that for each psychometric property, as well as development and content validity, there will be space to record the score for risk of bias (very good, adequate, doubtful, inadequate) following completion the COSMIN risk of bias checklist, and space to record quality of measurement property (sufficient, insufficient, indeterminate), following assessment using the COSMIN criteria for good measurement properties.

Table 2. Data extraction table, to completed for each measure, from the initial development article and any other relevant publications such as a manual.

Identifier		Measurement characteristics						
Name of measure	**Development article reference**	**Method of development** (e.g. items created from literature searching, previous measures, qualitative interviews)	**Intended context** (includes purpose [e.g. clinical /research?], setting, and method of administration)	**Study population for which it was developed** (language; cognitive ability; disease status; age etc)	**Availability in different settings / populations**	**N of items**	**N of scales / subscales**	**List of items & the domains / subscales they are reported to belong to**

Appendix D. Risk of bias checklist

Table 3. Boxes of the COSMIN Risk of Bias Checklist

Category	Boxes of the COSMIN Risk of Bias Checklist
Content Validity	Box 1. PROM development
	Box 2. Content validity
Internal Structure	Box 3. Structural validity
	Box 4. Internal consistency
	Box 5. Cross-cultural validity
Remaining measurement properties	Box 6. Reliability
	Box 7. Measurement error
	Box 8. Criterion validity
	Box 9. Hypothesis testing for construct validity
	Box 10. Responsiveness

Adapted with permission from Mokkink et al. (2017)

Each risk of bias checklist box is to be completed for each study that assesses that measurement property. Boxes 1 is to be completed for original development studies, whereas box 2 is to be completed for any additional content validity studies, or studies developing an established measure in a different population. Box 8 will not be completed for any study in this systematic review, as no gold standard measure of apathy exists. For details of how risk of bias is assessed for each measurement property, see Mokkink et al (2017).

Appendix E. COMSIN Quality Criteria

Table 4. COSMIN criteria for good measurement properties

Measurement property	Rating	Criteria
Structural Validity	+	**CTT:** CFA: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08[a] **IRT/Rasch:** No violation of unidimensionality[b]: CFI or TLI or comparable measure >0.95 OR RMSEA <0.06 OR SRMR <0.08 *AND* no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 *AND* no violation of monotonicity: adequate looking graphs OR item scalability >0.30 *AND* adequate model fit IRT: $\chi 2$ >0.01 Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > -2 and <2
	?	CTT: Not all information for '+' reported IRT/Rasch: Model fit not reported
	-	Criteria for '+' not met
Internal Consistency	+	At least low evidence[c] for sufficient structural validity[d] AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or Subscale[e]
	?	Criteria for "At least low evidence4 for sufficient structural Validity[d]" not met
	-	At least low evidence[c] for sufficient structural validity[5] AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale[e]
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC[d]
	?	MIC not defined
	-	SDC or LoA > MIC[d]
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis[f]
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis[f]
Cross-cultural validity \ measurement invariance	+	No important differences found between group factors (such as age, gender, language) in multiple group factor analysis OR no important DIF for group factors (McFadden's R^2 < 0.02)
	?	No multiple group factor analysis OR DIF analysis performed
	-	Important differences between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard < 0.70 OR AUC < 0.70

Responsiveness	+	The result is in accordance with the hypothesis[f] OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis[f] OR AUC < 0.70

AUC area under the curve, *CFA* confirmatory factor analysis, *CFI* comparative fit index, *CTT* classical test theory, *DIF* differential item functioning, *ICC* intraclass correlation coefficient, *IRT* item response theory, *LoA* limits of agreement, *MIC* minimal important change, *RMSEA* root mean square error of approximation, *SEM* standard error of measurement, *SDC* smallest detectable change, *SRMR* standardized root mean residuals, *TLI* Tucker–Lewis index

"+" = sufficient, "–" = insufficient, "?" = indeterminate

[a] To rate the quality of the summary score, the factor structures should be equal across studies

[b] unidimensionality refers to a factor analysis per subscale, while structural validity refers to a factor analysis of a (multidimensional) patient-reported outcome measure

[c] As defined by grading the evidence according to the GRADE approach

[d] This evidence may come from different studies

[e] The criteria 'Cronbach alpha < 0.95' was deleted, as this is relevant in the development phase of a PROM and not when evaluating an existing PROM.

[f] The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses

Table and footnotes reproduced with permission from Prinsen et al. (2018)

In addition criteria listed in this table, criteria previously described by Prinsen et al. (2016, p. 7) for studies using exploratory factor analysis will be applied: + "First factor accounts for at least 20% of the variability AND ratio of the variance explained by the first to the second factor greater than 4"; ? "Not all information for '+' reported"; - "Criteria for '+' not met".

Appendix F. Modified GRADE criteria

The modified GRADE criteria is described by Prinsen et al. (2018). Each measurement property (of each measure) starts with a 'high' quality evidence rating. The seriousness of each factor: risk of bias; inconsistency; imprecision; and indirectness, are then assessed. Where concerns about the quality of the evidence for the factor in question is 'not serious', no downgrading from 'high quality' (0) occurs. However where concern about the quality of evidence for a factor is serious, the evidence is downgraded to moderate quality (-1), where it is very serious, it is downgraded to low quality (-2), and where it is extremely serious, it is downgraded to very low quality (-3). Downgrading is additive, so that a judgement of 'serious' (-1) on two factors, and 'not serious' (0) on the other two factors, will mean that the overall quality of evidence for the measurement property of that measure is rated as 'low' (-2).

Risk of Bias

(0) *Not serious:* Multiple studies of at least adequate quality, OR one study of very good quality

(-1) *Serious:* Multiple studies of doubtful quality, OR one study of adequate quality

(-2) *Very serious:* Multiple studies of inadequate quality, OR one study of doubtful quality

(-3) *Extremely serious:* One study of inadequate quality

Inconsistency

(0) *Not serious:* Results of studies, or subgroups of studies with similar results, are pooled and summarized, with overall ratings using criteria for good measurement properties successfully provided.

(-1 or -2) *Serious or very serious:* Where unexplained inconsistencies exist, reviewers might pool or summarize the result based on the majority of results, or other appropriate judgements, and use the criteria for good measurement properties to rate the pooled results as sufficient (+) or insufficient (-). Whether the inconsistency is serious or very serious is context dependent and to be decided by the reviewers.

(n/a) Alternatively, when using the criteria for good measurement properties, results can be rated as 'inconsistent' (+/-). In this case, no quality of evidence rating is required.

Imprecision

This factor does not apply to measurement properties that have already taken into account the sample size. Content validity, structural validity, and cross-cultural validity are already used to

assess quality of evidence in the COSMIN Risk of Bias checklist, so should not be assessed for imprecision.

(0) Not serious: Total sample size of the pooled of summarized studies is 100 or above

(-1) Serious: Total sample size of the pooled or summarized studies is below 100

(-2) Very serious: Total sample size of the pooled or summarized studies is below 50

Indirectness

(0) Not serious: Studies sample characteristics matches the target population of the review. Construct validity and responsiveness is good.

(-1 or -2) Serious or very serious: Reasons for downgrading include: only part of the study population has characteristics relevant to the review; construct validity or responsiveness has weak evidence, for example when it is based on comparisons with measures of different constructs or on differences between participants with extremely different characteristics. Whether this is serious or very serious is context dependent and to be decided by the reviewers.