Reliability, validity and relevance of needs assessment instruments for informal dementia caregivers: a psychometric systematic review protocol

Stephanie Kipfer[1,4] · Manuela Eicher[2] · Annie Oulevey Bachmann[2,3,4] · Sandrine Pihet[1]

[1]School of Health Sciences Fribourg, University of Applied Sciences and Arts Western Switzerland, Fribourg, Switzerland, [2]Faculty of Biology and Medicine, Institute of Higher Education and Research in Healthcare (IFS), University of Lausanne, Lausanne, Switzerland, [3]School of Health Sciences Lausanne, University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland, and [4]Bureau d'Echange des Savoirs pour des pratiques exemplaires de soins (BEST): a Joanna Briggs Institute Centre of Excellence, Lausanne, Switzerland

Review question/objective: The objective of this psychometric review is to identify needs assessment instruments for informal dementia caregivers which are:

i. Reliable and valid in measuring the needs of informal dementia caregivers
ii. Relevant for clinical practice, research and informal caregivers

More specifically, the aim is to present an overview and an evaluation of the available needs assessment instruments, including: i) their psychometrics (reliability and validity) when available, and ii) their relevance according to the instrument characteristics, namely, their purpose, application method, administration burden, number of items and domain structure.

Keywords dementia; informal caregivers; instrument; needs assessment; psychometrics

JBI Database System Rev Implement Rep 2018; 16(2):269–286.

Introduction

Dementia is characterized by a progressive decline of cognitive and social functions. This limits the autonomy of those affected by dementia and makes it difficult for them to cope with daily life. They become increasingly dependent on the care of others, particularly, informal caregivers. Informal caregivers are individuals who regularly provide unpaid care, assistance and/or supervision to a close person with reduced autonomy.[1] In the EUROFAM-CARE study the informal caregivers reported memory problems in 46% of all caring situations and some behavioral problems in 34% of the care recipients.[2] Caring for someone with behavioral problems was stated as most burdensome for the informal caregivers. Studies show that informal dementia caregivers spend more time providing care over a longer time period than caregivers of people without dementia. The caregiving time increases even more if the person with dementia shows behavioral symptoms.[3] In the United States (US), for example, 31.1% of informal dementia caregivers provide care for two to three years, 18.5% for four to five years and 38.4% for six or more years.[4] In a survey conducted by Alzheimer Europe almost half of the informal dementia caregivers spent more than 10 hours per day providing care.[5] This is comparable to data from the US, where informal dementia caregivers spend on average nine hours per day caregiving.[6] In 2015 the estimated economic value of unpaid care provided by informal dementia caregivers in the US was $221.3 billion.[4] In Switzerland, as an example of a Western European country, informal caregivers contributed 64 million of unpaid hours, which saved the Swiss health care system 3.55 billion CHF, a 22% increase since 2010.[7] Informal caregivers contribution is expected to further increase due to the rising care needs of an aging population and the growing prevalence of (multiple) chronic conditions, in particular, dementia.[7] In addition the number of available formal carers is not expected to increase in line with the rising care needs of the aging population.[8] This shows that supporting informal caregivers and recognizing their valuable contribution is a significant public health issue.

Correspondence: Stephanie Kipfer, stephanie.kipfer@hefr.ch
There is no conflict of interest in this project.
DOI: 10.11124/JBISRIR-2017-003367

Informal caregivers play a key role not only for the people with dementia but also for society by promoting the sustainability of the health care system.[9-11]

Caring for a person with dementia is a challenging experience and the burden on informal dementia caregivers is higher in comparison to informal caregivers of persons with other chronic conditions.[12,13] Straining continuous care, an unpredictable course and neuropsychiatric symptoms of the person with dementia can cause high levels of stress which often lead to physical, psychological, emotional, social and financial problems.[3,13-15] In addition, family caregivers often have no experiences in performing care, feel unprepared and are lacking the required knowledge and support from health care providers to deliver appropriate care.[4,16] Informal caregivers report feelings of tiredness, stress, helplessness, and loneliness, and show a high prevalence of depression and anxiety.[3] Due to the nature of dementia, informal caregivers also struggle with feelings of guilt, ambivalence, grief and loss. Identified physical problems can create an increased risk for vascular disease, impaired wound healing, decreased immunity, and reduced likelihood to engage in preventive health behavior.[3,17] Poor physical and psychological health conditions not only impair the quality of life of informal dementia caregivers, but also affect the ability to provide care to the person with dementia, and to sustain the own social support network, which leads to social isolation.[3,17-19] Burden and health deterioration of informal dementia caregivers are core predictors of early institutionalization and mistreatment of their care recipient.[17,20]

Due to the challenges of caregiving and the associated burden, informal dementia caregivers report significant unmet needs at all stages of the disease. Their needs cover very diverse areas, such as information about the illness and support resources, support for own emotional concerns, support to communicate with the care recipient, the family or the service providers, practical support in daily care and respite, or financial support.[3,21-25]

Informal dementia caregivers often report that health care providers do not attend and adapt to their multiple needs sufficiently, resulting in care fragmentation and poor coordination, ultimately increasing stress and underutilization of support services despite the needs.[26-28] Underutilization of health care and other support resources contributes to the exhaustion of the informal dementia caregivers and precipitated institutionalization of their care recipient, and thereby increases health care costs.[20,29,30] Informal caregivers do not always spontaneously express how their needs can be met.[31] Therefore the evaluation of their needs by a professional is crucial in order that their needs can be met in a person centered way, the quality of life of the caregiver and the person affected can be enhanced, and the caring situation at home can be maintained.

Most studies on informal dementia caregivers needs have used qualitative study designs. Existing quantitative questionnaires have limitations: very few items for caregivers,[2,32] poor validation,[21,22,33,34] or lack of empirical evidence about need dimensions (factor structure). This limits their use in both research and clinical practice. In addition many of the assessment instruments, such as semi-structured interviews, are time-intensive (e.g. assessment alone lasts on average two hours[25]; or 90 minutes[22]). Furthermore, most of the collected information in interviews is qualitative, as such it is usually extensive, and more time is needed to prepare the information to make it available for the caregiver or other service providers (e.g. transcriptions). In view of the growing economic pressure on the social system and rising support needs associated with population aging, such resources are impossible to manage on a large scale.

A preliminary search in MEDLINE, CINAHL, *JBI Database of Systematic Reviews and Implementation Reports* and the Cochrane Database of Systematic Reviews was performed to identify completed and in progress systematic reviews on needs assessment instruments for informal dementia caregivers. To our knowledge there is no review which evaluates and compares the different needs assessment instruments for informal dementia caregivers according to their reliability, validity and relevance for clinical practice, research and informal caregivers. This information would be invaluable for researchers and clinicians to help them make informed decisions about the best instruments available for their specific purpose and clarify whether a new comprehensive needs assessment instrument is required.

Inclusion criteria

The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) guidelines for systematic reviews of measurement properties[38] recommend the following inclusion

criteria: i) the instrument should aim to measure the construct of interest (informal dementia caregiver needs), ii) the study sample should concern the target population of interest (informal dementia caregivers), iii) the study should concern the type of measurement instrument of interest (self-reported or professionally interviewed), iv) the aim of the study should be the development of a measurement instrument or the evaluation of one or more of its measurement properties (see types of studies). The development of the inclusion criteria for this review was guided by the COSMIN guidelines and the JBI template.

Participants

This review will consider studies that include informal caregivers of persons with dementia living at home. Informal caregivers are individuals who regularly provide unpaid care, assistance and/or supervision to a close person with reduced autonomy, in this context with dementia.[1]

Intervention

This review will consider studies that report on needs assessment instruments for assessing the needs of informal dementia caregivers. Needs can be defined as ''a condition that is important to the subject and that is not being satisfied in the subject's present environment''.[35(p.772)] The application method of the instruments can either be self-reported or professionally interviewed. For example, studies about the Carers' Needs Assessment for Dementia (CAN-D), the Care Needs Assessment Pack for Dementia (CARENAPD) or the John Hopkins Dementia Care Needs Assessment (JHDCNA).

Outcomes

This review will consider studies that include the following outcomes:

Psychometric properties:

- Reliability (test-retest reliability, inter-rater reliability, internal consistency)
- Validity (content validity, construct validity, structural validity, sensitivity to change, criterion-related validity).

Instrument characteristics:

- Purpose (original intended use)
- Application method (self-reported, professionally interviewed)
- Administration burden (training for clinicians, time for completion and evaluation for the clinicians and the informal dementia caregivers)
- Number of items and domain structure.

As not every article will provide data for all of these outcomes, articles which report at least on one outcome regarding reliability or validity will be considered. Criterion-related validity will only be considered if there is a reasonable gold-standard available (in accordance with COSMIN guidelines[38]).

Types of studies

This psychometric review will consider psychometric studies namely instrument development or instrument evaluation studies. Other types of studies (in which only needs assessment instruments are used) will be considered to identify eligible instruments and their responsible authors. Responsible authors will then be contacted in search for unpublished psychometric studies or testing of the instrument.

Methods

Search strategy

The search strategy aims to find both published and unpublished studies. A three-step search strategy will be utilized in this review. An initial limited search of MEDLINE and CINAHL will be undertaken followed by an analysis of the text words contained in the title and abstract, and of the index terms used to describe article. A second search using all identified keywords and index terms will then be undertaken across all included databases. Thirdly, the reference list of all identified reports and articles will be searched for additional studies. A complementary search will be performed in the included databases using the names of the needs assessment instruments identified in the three foregoing steps. Studies published in English, German and French will be considered for inclusion in this review. There will be no limitation regarding the publication time.

Information sources

The databases to be searched include: MEDLINE, OVID Nursing and PsycINFO via OVID, and CINAHL.

The search for unpublished studies will include: Google Scholar, ProQuest Dissertations and Theses, Researchgate (contact with relevant researchers), homepages with information about needs assessment/outcome tools and homepages of dementia associations or organizations.

We will also identify relevant researchers during the literature research and contact them by email to obtain information about unpublished psychometric studies or instrument testing of relevant instruments. The number of articles retrieved through this approach and the process itself will be described in the section ''Search strategy'' of the review report.

Initial keywords to be used will be: dementia; informal caregivers; caregiver needs; needs assessment; needs evaluation; instrument; interview; assessment; questionnaire; validation; psychometrics; reliability; validity.

Assessment of methodological quality

Papers selected for retrieval will be assessed by two independent reviewers for methodological validity prior to inclusion in the review. To assess the methodological quality the COSMIN checklist will be used (Appendix I). The COSMIN checklist is a standardized tool which is recommended to use in systematic reviews of measurement properties.[36] This tool meets the specific needs of a psychometric review and has been suggested before in another Johanna Briggs Institute review protocol.[37] It will therefore be preferable over the standardized critical appraisal instruments from the Joanna Briggs Institute. The checklist consists of 12 different boxes with five to 18 items per box. Seven of these boxes consider psychometric properties, namely, internal consistency, reliability, measurement error, content validity, structural validity, criterion-related validity and responsiveness. The general methodological quality will be assessed with the box, hypotheses testing. There is also one box for additional methodological standards for studies using item response theory models, one for studies investigating interpretability and one for cross-cultural validity. These three boxes will not be used in this review as they focus on more advanced properties. The remaining box concerning generalizability is recommended to use for data extraction. The COSMIN checklist is a modular tool. The selection of boxes to be used is determined by the measurement properties evaluated in the reviewed study. The checklist contains a four-point response option to evaluate the different items: excellent, good, fair and poor. However some items have only two or three response options. A methodological quality score per box can be obtained by taking the lowest rating of any item in the box (''worst score count''). The lowest score of any box presents the overall score of the reviewed study. Studies with poor scores in all boxes will be excluded from the review. Any disagreements that arise between the reviewers will be resolved through discussion, or with a third reviewer.

Data extraction

Data will be extracted from papers included in the review using a standardized data extraction tool developed for this review (Appendix II). This tool is inspired by the different elements of the standardized data extraction tools from Joanna Briggs Institute System for the Unified Management, Assessment and Review of Information (JBI SUMARI)and adapted to the specific elements of a psychometric review. The data extracted will include specific details about: i) the study characteristics, ii) the instrument characteristics, and iii) the outcomes of significance to the review question and specific objectives.

Study characteristics: citation details, aim of the study, study design and method, setting, population characteristics, definition of informal caregivers and needs.

Instrument characteristics: name of the instrument, purpose, target population, application method, respondent and/or administrative burden, number of items and domain structure, range of scores, response options/format.

Psychometric outcomes: reliability (test-retest reliability, inter-rater reliability, internal consistency), validity (content validity, construct validity, structural validity, criterion-related validity, sensitivity to change).

As mentioned above, the box concerning generalizability of the COSMIN checklist will also be used. Data will be extracted independently by two reviewers. Disagreement between the reviewers will be resolved by discussion, or with a third reviewer. If necessary, the responsible researchers of an instrument will be contacted for missing or additional data.

Data synthesis

The main aim of the data synthesis is to compare outcomes to provide recommendations on the most suitable instrument for research, clinical use and informal caregivers. The findings about reliability, validity and the instrument characteristics will be

compared and presented in narrative form including tables and figures to aid data presentation. A content comparison will give an overview of the content of each instrument and the similarities and differences on an item level.[38] However, if we can identify substantially more than one published or unpublished psychometric study per instrument, summary measures will be computed for this instrument, for example, the average Cronbach's alpha with its 95% confidence interval. The information gathered in the generalization box can be used to assess the similarity of the participants in these studies.[38] To judge the psychometric outcomes of the different instruments, the quality criteria from Terwee *et al.* will be used.[39] These criteria will allow us to judge: content validity, internal consistency, criterion-related validity, construct validity, test-retest reliability (agreement), inter-rater reliability (reliability) and sensitivity to change (responsiveness) in terms of positive rating, indeterminate rating, negative rating, no information available and doubtful design or method (Appendix III). Floor and ceiling effects and interpretability will not be rated as there is no gold-standard. Only the most common and basic properties will be evaluated in this review. For structural validity a quality criteria will be developed if necessary. The results of this appraisal will be presented in a narrative form.

Acknowledgements

We thank Prof. Dr. Dawn Carnes (School of Health Sciences Fribourg, University of Applied Sciences and Arts Western Switzerland) for comments that greatly improved the protocol.

References

1. Swiss Confederation. Support for family caregivers: An analysis of the situation and the required measures in Switzerland [Soutien aux proches aidants: Analyse de la situation et mesures requises pour la Suisse]. Bern: Federal Council; 2014.
2. EUROFAMCARE. EUROFAMCARE - Services for Supporting Family Carers of Older Dependent People in Europe: Characteristics, Coverage and Usage Brussels: EUROFAMCARE Consortium; 2006.
3. Thompson GN, Roger K. Understanding the needs of family caregivers of older adults dying with dementia. Palliat Support Care 2014;12(3):223–31.
4. Alzheimer's Association. 2016 Alzheimer's Disease Facts and Figures: Caregiving. Alzheimer's & Dementia 2016;12(4):31–4.
5. Alzheimer Europe. Who cares? The state of dementia care in Europe. Luxembourg: Alzheimer Europe; 2006.
6. Fisher GG, Franks MM, Plassman BL, Brown SL, Potter GG, Llewellyn D, *et al*. Caring for individuals with dementia and cognitive impairment, not dementia: findings from the aging, demographics, and memory study. J Am Geriatr Soc 2011;59(3):488–94.
7. De Pietro C, Camenzind P, Sturny I, Crivelli L, Edwards-Garavoglia S, Spranger A, *et al*. Switzerland - Health system review. Brussels: European Observatory on Health Systems and Policies; 2015.
8. Zwaanswijk M, Peeters JM, van Beek AP, Meerveld JH, Francke AL. Informal caregivers of people with dementia: problems, needs and support in the initial stage and in subsequent stages of dementia: a questionnaire survey. Open Nurs J 2013;7:6–13.
9. Kesselring A. Caring for family members at home [Angehörige zu Hause pflegen: Anatomie einer Arbeit]. Swiss Medical Journal [Schweizerische Ärztezeitung] 215 2004; 85(10):504–6.
10. Rosa E, Lussignoli G, Sabbatini F, Chiappa A, Di Cesare S, Lamanna L, *et al*. Needs of caregivers of the patients with dementia. Arch Gerontol Geriatr 2010;51(1):54–8.
11. Kraft E, Marti M, Werner S, Sommer H. Cost of dementia in Switzerland. Swiss Med Wkly 2010;140:w13093.
12. Butcher HK, Holkup PA, Buckwalter KC. The experience of caring for a family member with Alzheimer's disease. West J Nursing Res 2001;23(1):33–55.
13. de la Cuesta C. The craft of care: family care of relatives with advanced dementia. Qual Health Res 2005;15(7):881–96.
14. Vellone E, Sansoni J, Cohen MZ. The experience of Italians caring for family members with Alzheimer's disease. J Nurs Scholarsh 2002;34(4):323–9.
15. Knapp M, Prince M, Albanese E, Banerjee S, Dhanasiri S, Fernandez J-L, *et al*. Dementia UK. London: Alzheimer's Society; 2007.
16. Reinhard SC, Given B, Petlick NH, Bemis A. Supporting family caregivers in providing care. In: Hughes RG, editor. Patient safety and quality: An evidence-based handbook for nurses. Rockville (MD): Agency for Healthcare Research and Quality (US), 2008;341–404.
17. Schulz R, Martire LM. Family caregiving of persons with dementia: prevalence, health effects, and support strategies. Am J Geriatr psychiatry 2004;12(3):240–9.
18. Perrig-Chiello P, Höpflinger F, Schnegg B. SwissAgeCare: Nursing relatives of older people in the Switzerland Bern: Universität Bern; 2010.
19. Brodaty H, Donkin M. Family caregivers of people with dementia. Dialogues Clin Neurosci 2009;11(2):217–28.
20. Gaugler JE, Kane RL, Kane RA, Clay T, Newcomer R. Caregiving and institutionalization of cognitively impaired older people: utilizing dynamic predictors of change. Gerontologist 2003;43(2):219–29.
21. Black BS, Johnston D, Rabins PV, Morrison A, Lyketsos C, Samus QM. Unmet needs of community-residing persons

with dementia and their informal caregivers: findings from the maximizing independence at home study. J Am Geriatr Soc 2013;61(12):2087–95.
22. Ducharme F, Kergoat MJ, Coulombe R, Levesque L, Antoine P, Pasquier F. Unmet support needs of early-onset dementia family caregivers: a mixed-design study. BMC Nurs 2014;13(1):49.
23. Bass DM, Judge KS, Snow AL, Wilson NL, Morgan R, Looman WJ, *et al*. Caregiver outcomes of partners in dementia care: effect of a care coordination program for veterans with dementia and their family members and friends. J Am Geriatr Soc 2013;61(8):1377–86.
24. Afram B, Verbeek H, Bleijlevens MH, Hamers JP. Needs of informal caregivers during transition from home towards institutional care in dementia: a systematic review of qualitative studies. Int psychogeriatr 2015;27(6):891–902.
25. Freudiger Pittet S, Jordan A. Evaluation de la charge et des besoins des proches aidants. Lausanne: Association vaudoise d'aide et de soins à domicile (AVASD); 2012.
26. Brodaty H, Thomson C, Thompson C, Fine M. Why caregivers of people with dementia and memory loss don't use services. Int J Geriatr Psychiatry 2005;20(6):537–46.
27. Bass DM, Judge KS, Snow AL, Wilson NL, Looman WJ, McCarthy C, *et al*. Negative caregiving effects among caregivers of veterans with dementia. Am J geriatr psychiatry 2012;20(3):239–47.
28. Karlsson S, Bleijlevens M, Roe B, Saks K, Martin MS, Stephan A, *et al*. Dementia care in European countries, from the perspective of people with dementia and their caregivers. J Adv Nurs 2015;71(6):1405–16.
29. Gaugler JE, Kane RL, Kane RA, Newcomer R. Unmet care needs and key outcomes in dementia. J Am Geriatr Soc 2005;53(12):2098–105.
30. Gaugler JE, Kane RL, Kane RA, Newcomer R. Early community-based service utilization and its effects on institutionalization in dementia caregiving. Gerontologist 2005;45(2):177–85.
31. van der Roest HG, Meiland FJ, Maroccini R, Comijs HC, Jonker C, Droes RM. Subjective needs of people with dementia: a review of the literature. Int psychogeriatr 2007;19(3):559–92.
32. Reynolds T, Thornicroft G, Abas M, Woods B, Hoe J, Leese M, *et al*. Camberwell Assessment of Need for the Elderly (CANE): Development, validity and reliability. BJPsych 2000; 176(5):444–52.
33. Wancata J, Krautgartner M, Berner J, Alexandrowicz R, Unger A, Kaiser G, *et al*. The Carers' Needs Assessment for Dementia (CNA-D): development, validity and reliability. Int psychogeriatr 2005;17(3):393–406.
34. McWalter G, Toner H, McWalter A, Eastwood J, Marshall M, Turvey T. A community needs assessment: the care needs assessment pack for dementia (CarenapD)–its development, reliability and validity. Int J Geriatr Psychiatry 1998; 13(1):16–22.
35. Hileman JW, Lackey NR, Hassanein RS. Identifying the needs of home caregivers of patients with cancer. Oncol nurs forum 1992;19(5):771–7.
36. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21(4):651–7.
37. Simpelaere I, White A, Bekkering GE, Geurden B, Van Nuffelen G, De Bodt M. Patient-reported and proxy-reported outcome measures for the assessment of health-related quality of life among patients receiving enteral feeding: a systematic review protocol. JBI database of systematic reviews and implementation reports 2016;14(7):45–75.
38. Terwee CB. Protocol for systematic reviews of measurement properties. Amsterdam: VU University Medical Center Knowledgecenter Measurement Instruments; 2011.
39. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60(1):34–42.
40. Terwee CB. COSMIN checklist with 4-point scale. Amsterdam: University Medical Center, Department of Epidemiology and Biostatics, EMGO Institute for Health and Care Research; 2011.
41. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, *et al*. COSMIN checklist manual. Amsterdam: University Medical Center, Department of Epidemiology and Biostatics, EMGO Institute for Health and Care Research; 2012

Appendix I: Appraisal instrument

Adapted from the COSMIN checklist[40]

Box internal consistency	excellent	good	fair	poor
1. Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4. Was the sample size included in the internal consistency analysis adequate?	Adequate sample size (≥100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
5. Was the unidimensionality of the scale checked? i.e. was factor analysis or item response theory (IRT) model applied?	Factor analysis performed in the study population	Authors refer to another study in which factor analysis was performed in a similar study population	Authors refer to another study in which factor analysis was performed, but not in a similar study population	Factor analysis NOT performed and no reference to another study
6. Was the sample size included in the unidimensionality analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 6–7* #items but <100	5* #items but <100	<5* #items
7. Was an internal consistency statistic calculated for each (unidimensional) (sub)-scale separately?	Internal consistency statistic calculated for each subscale separately			Internal consistency statistic NOT calculated for each subscale separately
8. Were there any important flaws in the design or methods of the study?	No other important methodo-logical flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study	Other important methodo-logical flaws in the design or execution of the study
Statistical methods				
9. for classical test theory (CTT), continuous scores: Was Cronbach's alpha calculated?	Cronbach's alpha calculated		Only item-total correlations calculated	No Cronbach's alpha and no item-total correlations calculated
10. for CTT, dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	Cronbach's alpha or KR-20 calculated		Only item-total correlations calculated	No Cronbach's alpha or KR-20 and no item-total correlations calculated
11. for IRT: Was a goodness of fit statistic at a global level calculated? E.g. $\chi 2$, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	Goodness of fit statistic at a global level calculated			Goodness of fit statistic at a global level NOT calculated

NB. Item 1 is used to determine whether internal consistency is relevant for the instrument under study. It is not used to rate the quality of the study

Box reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)	excellent	good	fair	poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
4. Were at least two measurements available?	At least two measurements			Only one measurement
5. Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measure-ments NOT independent
6. Was the time interval stated?	Time interval stated		Time interval NOT stated	
7. Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8. Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9. Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10. Were there any important flaws in the design or methods of the study?	No other important methodo-logical flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study	Other important methodo-logical flaws in the design or execution of the study
Statistical methods				
11. for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated
12. for dichotomous/nominal/ ordinal scores: Was kappa calculated?	Kappa calculated			Only percentage agreement calculated
13. for ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated	Only percentage agreement calculated
14. for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	Weighting scheme described	Weighting scheme NOT described		

Box measurement error: absolute measures	excellent	good	fair	poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
4. Were at least two measurements available?	At least two measurements			Only one measurement
5. Were the administrations independent?	Independent measurements	Assumable that the measurements were independent	Doubtful whether the measurements were independent	Measurements NOT independent
6. Was the time interval stated?	Time interval stated		Time interval NOT stated	
7. Were patients stable in the interim period on the construct to be measured?	Patients were stable (evidence provided)	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable
8. Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate	Time interval NOT appropriate
9. Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar
10. Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodological flaws in the design or execution of the study	Other important methodological flaws in the design or execution of the study
Statistical methods				
11. for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	SEM, SDC, or LoA calculated	Possible to calculate LoA from the data presented		SEM calculated based on Cronbach's alpha, or on SD from another population

Box content validity (including face validity)	excellent	good	fair	poor
General requirements				
1. Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	Assessed if all items refer to relevant aspects of the construct to be measured		Aspects of the construct to be measured poorly described AND this was not taken into consideration	NOT assessed if all items refer to relevant aspects of the construct to be measured
2. Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	Assessed if all items are relevant for the study population in adequate sample size ($\geq$10)	Assessed if all items are relevant for the study population in moderate sample size (5–9)	Assessed if all items are relevant for the study population in small sample size ($<$5)	NOT assessed if all items are relevant for the study population OR target population not involved
3. Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	Assessed if all items are relevant for the purpose of the application	Purpose of the instrument was not described but assumed	NOT assessed if all items are relevant for the purpose of the application	
4. Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	Assessed if all items together comprehensively reflect the construct to be measured		No theoretical foundation of the construct and this was not taken into consideration	NOT assessed if all items together comprehen-sively reflect the construct to be measured
5. Were there any important flaws in the design or methods of the study?	No other important methodological flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study	Other important methodo-logical flaws in the design or execution of the study

Box structural validity	excellent	good	fair	poor
1. Does the scale consist of effect indicators, i.e. is it based on a reflective model?				
Design requirements				
2. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
3. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
4. Was the sample size included in the analysis adequate?	7* #items and ≥100	5* #items and ≥100 OR 5–7* #items but <100	5* #items but <100	<5* #items
5. Were there any important flaws in the design or methods of the study?	No other important methodo-logical flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study (e.g. rotation method not described)	Other important methodo-logical flaws in the design or execution of the study (e.g. inappropriate rotation method)
Statistical methods				
6. for CTT: Was exploratory or confirmatory factor analysis performed?	Exploratory or confirmatory factor analysis performed and type of factor analysis appropriate in view of existing information	Exploratory factor analysis performed while confirmatory would have been more appropriate		No exploratory or confirmatory factor analysis performed
7. for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	IRT test for determining (uni)di-mensionality performed			IRT test for determining (uni)dimensionality NOT performed

Box hypotheses testing	excellent	good	fair	poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size (≥100 per analysis)	Good sample size (50–99 per analysis)	Moderate sample size (30–49 per analysis)	Small sample size (<30 per analysis)
4. Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Multiple hypotheses formulated a priori	Minimal number of hypotheses formulate a priori	Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
5. Was the expected direction of correlations or mean differences included in the hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
6. Was the expected absolute or relative magnitude of correlations or mean differences included in the hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
7. for convergent validity: Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)	Adequate description of most of the constructs measured by the comparator instrument(s)	Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
8. for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	No information on the measurement properties of the comparator instrument(s)
9. Were there any important flaws in the design or methods of the study?	No other important methodo-logical flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodo-logical flaws in the design or execution of the study
Statistical methods				
10. Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate	Assumable that statistical methods were appropriate, e.g. Pearson correlations applied, but distribution of scores or mean (SD) not presented	Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Box Criterion-related validity	excellent	good	fair	poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
4. Can the criterion used or employed be considered as a reasonable 'gold standard'?	Criterion used can be considered an adequate 'gold standard' (evidence provided)	No evidence provided, but assumable that the criterion used can be considered an adequate 'gold standard'	Unclear whether the criterion used can be considered an adequate 'gold standard'	Criterion used can NOT be considered an adequate 'gold standard'
5. Were there any important flaws in the design or methods of the study?	No other important methodologi-cal flaws in the design or execution of the study		Other minor methodologi-cal flaws in the design or execution of the study	Other important methodologi-cal flaws in the design or execution of the study
Statistical methods				
6. for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	Correlations or AUC calculated			Correlations or AUC NOT calculated
7. for dichotomous scores: Were sensitivity and specificity determined?	Sensitivity and specificity calculated			Sensitivity and specificity NOT calculated

Box responsiveness	excellent	good	fair	poor
Design requirements				
1. Was the percentage of missing items given?	Percentage of missing items described	Percentage of missing items NOT described		
2. Was there a description of how missing items were handled?	Described how missing items were handled	Not described but it can be deduced how missing items were handled	Not clear how missing items were handled	
3. Was the sample size included in the analysis adequate?	Adequate sample size (≥100)	Good sample size (50–99)	Moderate sample size (30–49)	Small sample size (<30)
4. Was a longitudinal design with at least two measurement used?	Longitudinal design used			No longitudinal design used
5. Was the time interval stated?	Time interval adequately described			Time interval NOT described
6. If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	Anything that occurred during the interim period (e.g. treatment) adequately described	Assumable what occurred during the interim period	Unclear or NOT described what occurred during the interim period	
7. Was a proportion of the patients changed (i.e. improvement or deterioration)?	Part of the patients were changed (evidence provided)	NO evidence provided, but assumable that part of the patients were changed	Unclear if part of the patients were changed	Patients were NOT changed

(*Continued*)

Box responsiveness	excellent	good	fair	poor
Design requirements for hypotheses testing For constructs for which a gold standard was not available:				
8. Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	Hypotheses formulated a priori		Hypotheses vague or not formulated but possible to deduce what was expected	Unclear what was expected
9. Was the expected direction of correlations or mean differences of the change scores of the instruments included in these hypotheses?	Expected direction of the correlations or differences stated	Expected direction of the correlations or differences NOT stated		
10. Were the expected absolute or relative magnitude of correlations or mean differences of the change scores of the instruments included in these hypotheses?	Expected magnitude of the correlations or differences stated	Expected magnitude of the correlations or differences NOT stated		
11. Was an adequate description provided of the comparator instrument(s)?	Adequate description of the constructs measured by the comparator instrument(s)		Poor description of the constructs measured by the comparator instrument(s)	NO description of the constructs measured by the comparator instrument(s)
12. Were the measurement properties of the comparator instrument(s) adequately described?	Adequate measurement properties of the comparator instrument(s) in a population similar to the study population	Adequate measurement properties of the comparator instrument(s) but not sure if these apply to the study population	Some information on measurement properties (or a reference to a study on measurement properties) of the comparator instrument(s) in any study population	NO information on the measurement properties of the comparator instrument(s)
13. Were there any important flaws in the design or methods of the study?	No other important methodo-logical flaws in the design or execution of the study		Other minor methodo-logical flaws in the design or execution of the study (e.g. only data presented on a comparison with an instrument that measures another construct)	Other important methodo-logical flaws in the design or execution of the study
Statistical methods				
14. Were design and statistical methods adequate for the hypotheses to be tested?	Statistical methods applied appropriate		Statistical methods applied NOT optimal	Statistical methods applied NOT appropriate

Appendix II: Data extraction instrument

Reviewer	
Date	
Record Number	

Study characteristics	
Title	
Authors	
Year of publication	
Journal	
Aim of the study	
Study design/method	
Setting	
Population characteristics	
Definition of informal caregiver's needs	
Instrument characteristics	
Instrument name	
Purpose	
Target population	
Application method	
Respondent administrative burden	
Domain structure	
Number of items (total/per domain)	
Range of scores	
Response options/format	
Outcomes	
Reliability	
Test-retest reliability	
Inter-rater reliability	
Internal consistency	

Validity	
Content validity	
Construct validity	
Structural validity	
Criterion-related validity	
Sensitivity to change	

Box Generalizability[40,41]

Median or mean age (with standard deviation or range)	
Distribution of sex	
Important disease characteristics (e.g. severity, status, duration) and description of treatment	
Setting(s) in which the study was conducted	
Countries in which the study was conducted	
Languages in which the instrument was evaluated	
Method used to select patients	
Percentage of missing responses (response rate)	

Results	
Authors conclusions	
Reviewers conclusions	

Appendix III: Quality criteria for measurement properties of questionnaires

Adapted version, for complete version see Terwee et al.[39]

Property	Definition	Quality criteria
1. Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire	+ A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection AND target population and (investigators OR experts) were involved in item selection; ? A clear description of above-mentioned aspects is lacking OR only target population involved OR doubtful design or method; − No target population involvement; 0 No information found on target population involvement.
2. Internal consistency	The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct	+ Factor analyses performed on adequate sample size (7 * # items and ≥100) AND Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95; ? No factor analysis OR doubtful design or method; − Cronbach's alpha(s) < 0.70 or > 0.95, despite adequate design and method; 0 No information found on internal consistency.
3. Criterion-related validity	The extent to which scores on a particular questionnaire relate to a gold standard	+ Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥ 0.70; ? No convincing arguments that gold standard is "gold" OR doubtful design or method; − Correlation with gold standard < 0.70, despite adequate design and method; 0 No information found on criterion-related validity.
4. Construct validity	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured	+ Specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses; ? Doubtful design or method (e.g., no hypotheses); − Less than 75% of hypotheses were confirmed, despite adequate design and methods; 0 No information found on construct validity.
5. Reproducibility **5.1 Agreement** (Test-retest reliability)	The extent to which the scores on repeated measures are close to each other (absolute measurement error)	+ MIC < SDC OR MIC outside the LOA OR convincing arguments that agreement is acceptable; ? Doubtful design or method OR (MIC not defined AND no convincing arguments that agreement is acceptable); − MIC ≥ SDC OR MIC equals or inside LOA, despite adequate design and method; 0 No information found on agreement.
5.2 Reliability (Inter-rater reliability)	The extent to which patients can be distinguished from each other, despite measurement errors (relative measurement error)	+ ICC or weighted Kappa ≥ 0.70; ? Doubtful design or method (e.g., time interval not mentioned); − ICC or weighted Kappa < 0.70, despite adequate design and method; 0 No information found on reliability.

(*Continued*)

Property	Definition	Quality criteria
6. Responsiveness (Sensitivity to change)	The ability of a questionnaire to detect clinically important changes over time	+ SDC or SDC < MIC OR MIC outside the LOA OR RR > 1.96 OR AUC ≥ 0.70; ? Doubtful design or method; − SDC or SDC ≥ MIC OR MIC equals or inside LOA OR RR ≤ 1.96 OR AUC < 0.70, despite adequate design and methods; 0 No information found on responsiveness.

ICC, Intraclass correlation; LOA, limits of agreement; MIC, minimal important change; SD, standard deviation; SDC, smallest detectable change.
+ = positive rating.
? = indeterminate rating.
− = negative rating.
0 = no information available.
Doubtful design or method = lacking of a clear description of the design or methods of the study, sample size smaller than 50 subjects (should be at least 50 in every (subgroup) analysis), or any important methodological weakness in the design or execution of the study.

