

Apathy Measures in Older Adults and People with Dementia: A Systematic Review of Measurement Properties using the COSMIN methodology

Clare Burgon^{a,b}, Sarah Elizabeth Goldberg^a, Veronika van der Wardt^{b,c}, Catherine Brewin^{a,d}, Rowan H Harwood^{a,d}

^a School of Health Sciences, University of Nottingham, Nottingham, United Kingdom.

^b Division of Rehabilitation, Ageing and Wellbeing, School of Medicine, University of Nottingham, Nottingham, United Kingdom.

^c Department of General Medicine, Preventative and Rehabilitative Medicine, Philipps-Universität Marburg, Marburg, Germany

^d Health Care of Older People, Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom.

Short Title: Systematic review of apathy measures in older adults and dementia

Corresponding Author:

Rowan H Harwood

School of Health Sciences

University of Nottingham, Queen's Medical Centre

Nottingham, Nottinghamshire, NG7 2HA, United Kingdom

Tel: 0115 823 0873

E-mail: rowan.harwood@nottingham.ac.uk

Number of Tables: 1

Number of Figures: 1

Word count: 5242

Keywords: Apathy; Measurement; Scale; Systematic Review; COSMIN.

1 **Abstract**

2 **Background**

3 Apathy is highly prevalent in dementia and is also seen in mild cognitive impairment and the general
4 population. Apathy contributes to failure to undertake daily activities, and can lead to health
5 problems or crises. It is therefore important to assess apathy. However, there is currently no gold-
6 standard measure of apathy. A comprehensive systematic review of the measurement properties of
7 apathy scales is required.

8 **Methods**

9 A systematic review was registered with PROSPERO (ID: CRD42018094390). MEDLINE, EMBASE,
10 PsycINFO and CINAHL were searched for studies that aimed to develop or assess the validity or
11 reliability of an apathy scale in participants over 65 years, living in the community. A systematic
12 review was conducted in line with the COnsensus-based Standards for the selection of health
13 Measurement INstruments (COSMIN) procedure for reviewing patient-reported outcome measures.
14 The studies' risk of bias were assessed and all relevant measurement properties were assessed for
15 quality. Results were pooled and rated using a modified Grading of Recommendations Assessment,
16 Development and Evaluation procedure.

17 **Results**

18 Fifty-seven publications regarding 18 measures and 39 variations met the eligibility criteria. The
19 methodological quality of individual studies ranged from inadequate to very good and measurement
20 properties ranged from insufficient to sufficient. Similarly, the overall evidence for measurement
21 properties ranged from very low to high quality. The Apathy Evaluation Scale and Lille Apathy Rating
22 Scale had sufficient content validity, reliability, construct validity, and where applicable, structural
23 validity and internal consistency.

24 **Conclusion**

25 Numerous scales are available to assess apathy, with varying psychometric properties. The Apathy
26 Evaluation Scale and Lille Apathy Rating Scale are recommended for measuring apathy in older adults
27 and people living with dementia. The apathy dimension of the commonly-used Neuropsychiatric
28 Inventory should be limited to screening for apathy.

29 **Introduction**

30 Apathy is a multidimensional construct, defined as quantitatively reduced behavioural, cognitive,
31 emotional or social (goal-directed) activity which may include reduced motivation, initiative, effort,
32 interest, concern about self or others, and affect [1]. Apathy is the most common neuropsychiatric
33 symptom of dementia [2] and is reported in 15% to 92% of people with dementia [3], and 12% to
34 40% of people with Mild Cognitive Impairment (MCI) [4,5]. Apathy is associated with important
35 outcomes in dementia and MCI, including disabilities in everyday functioning [6], increased carer
36 burden [7–10], worse adherence to interventions [11,12], and worse quality of life [13]. Prevention
37 or management of apathy in dementia has been identified as a priority area for research [14]. Apathy
38 in older adults is associated with increased likelihood of subsequent cognitive impairment [15],
39 conversion from MCI to dementia [5,16], and worse disease progression [17,18]. It is therefore
40 important to research across the spectrum of cognitive impairment [19], including older adults who
41 otherwise appear cognitively unimpaired.

42 There is no gold-standard measure of apathy [20]. Two systematic reviews of apathy scales have
43 been published; the first examined scales developed for people with neurodegenerative conditions
44 such as Parkinson’s Disease, Amyotrophic Lateral Sclerosis, and dementia [21], whilst the second
45 examined evidence for measurement properties in people with dementia [22]. The first systematic
46 review only included studies that assessed both validity and reliability of a scale within a single
47 publication. Whilst a scale should be both valid and reliable, separately published studies of reliability
48 and validity can collectively offer sufficient evidence for both. The latter review used limited search
49 criteria and it is unclear when the search was conducted. Therefore, important studies regarding the
50 quality of apathy scales may have been missed.

51 Both of these systematic reviews used the Quality Assessment of Diagnostic Accuracy Studies
52 (QUADAS) tool, designed for studies of diagnostic accuracy [23], not other measurement properties,
53 such as reliability. It is not clear how these reviews rated properties such as reliability using the
54 QUADAS criteria that refer to a ‘reference standard’, which is only relevant to properties such as
55 criterion validity. The COnsensus-based Standards for the selection of health Measurement
56 INstruments (COSMIN) programme of work has since published guidance for conducting and
57 reporting systematic reviews of health measures, with methodological quality standards and
58 measurement property quality criteria that enables a systematic and standardized critical
59 examination of all key measurement properties of scales [24,25].

60 The aim of this study was to assess and compare the quality of measurement properties (i.e. the
61 reliability and validity) and characteristics, of apathy scales and to analyse the quality of the evidence
62 in healthy older adults and people with dementia, in accordance with COSMIN guidance.

63 **Methods**

64 **Design**

65 This systematic review protocol was registered with PROSPERO (ID: CRD42018094390) and published
66 prior to analysis [26]. The COSMIN methodology for conducting systematic reviews of patient
67 reported outcome measures [24,25,27] was followed. Properties were assessed in relation to both
68 people with dementia and older adults. Some additional decisions were required for this review,
69 which were based on literature, discussions with the review team, and PPI input. For example, there
70 is no gold standard time interval for test-retest reliability studies [28] though a time interval of two
71 weeks is common [29]. Apathy is a relatively stable, but not an enduring trait, so a time interval that
72 exceeded 28 days or 1 month was considered inappropriate. A time interval of less than 3 days for
73 people with memory problems, and less than 7 days for people without memory problems, was also
74 considered inappropriate as memory of previous answers may inflate reliability estimates. These
75 were arbitrary numbers chosen in lieu of guidance, but was deemed acceptable by the review team
76 and PPI members.

77 **Searching, screening and selection**

78 MEDLINE (In-Process, Other Non-Indexed Citations and 1946 onwards) EMBASE (1980 onwards),
79 PsycINFO (1806 onwards, via Ovid) and CINAHL (1937 onwards) were searched using the specified
80 search strategy on 18th April 2018, and the search was re-ran in the same databases on 6th May 2020.
81 The reference lists of the included studies, and of any relevant review articles, were also examined
82 for relevant publications. The COSMIN search strategy for identifying research on the development,
83 validity or reliability of health related outcome measures [30] formed part of the search strategy,
84 along with apathy related terms (e.g. apathy; lack of or diminished motivation, interest, initiative;
85 emotional blunting; emotional responsiveness; abulia; anhedonia; frontal symptom; asocial; avolition;
86 lassitude). The search strategy was first created for MEDLINE (Supplementary Additional File 1), then
87 the subject headings and syntax were adapted for the other databases.

88 Inclusion criteria: studies that aimed to develop or assess the measurement properties of an apathy
89 scale based on patient or informant reports or interviewer or clinician ratings; primary research; full-

90 text publication; majority of participants living in the community; majority of participants aged 65 or
91 over. Exclusion criteria: studies assessing scales of apathy in a specific context such as work
92 performance, or following an event, such as in post-traumatic stress disorder and post-natal
93 depression. Additionally, development studies were included regardless of eligibility criteria if they
94 pertained to an apathy scale that was included in one of the eligible studies.

95 The titles and abstracts of articles were screened (by ClB) to assess whether they met the eligibility
96 criteria. All included full-text articles were assessed against eligibility criteria (by ClB), and a randomly
97 selected 10% of articles was independently assessed by a second reviewer (CaB). Articles for which
98 there was disagreement between reviewers were discussed and an agreement was reached.

99 Data extraction and assessment

100 Data extraction was conducted (by ClB) into a data extraction table (Supplementary Additional File
101 2). Data extraction of 20% of publications was checked by second reviewers (CaB; VvdW) and no
102 errors were found. For each study included in the review, data relating to study characteristics and
103 methods, participant characteristics, and measurement characteristics and properties were
104 extracted. Measurement properties included that of reliability (internal consistency, measurement
105 error and test-retest and interrater reliability) and validity (content validity, structural validity,
106 hypothesis testing for construct validity), as defined by the COSMIN taxonomy [31]. Criterion validity
107 and responsiveness were not reviewed, as there is no gold standard measure of apathy against which
108 to assess the scales.

109 Risk of bias in individual studies was examined using the COSMIN risk of bias checklist [25,27]
110 (Supplementary Additional File 3). The results of studies were assessed using COSMIN criteria for
111 good measurement properties [24,25]. Publications were assigned a number and randomly selected
112 for second reviewer ratings using a random number function in Microsoft Excel. Twenty-one percent
113 (N=12) of publications were independently rated by second reviewers for risk of bias and against
114 criteria for good measurement properties (SG and VvdW). Where there was disagreement, this was
115 discussed between the two raters (ClB and SG; ClB and VvdW) and any remaining disagreements
116 were discussed with a third rater.

117 Synthesis of results

118 Studies meeting the eligibility criteria were summarised using a narrative synthesis. For each scale,
119 the measurement properties reported in the corresponding studies were summarised, and the

120 quality of these synthesized results was assessed using the criteria for good measurement properties
121 [24,32]. Where there were different versions of scale, results were pooled providing they were not
122 contradictory. The COSMIN modified GRADE approach [24,25] was used to assess the quality of the
123 cumulative evidence for each measurement property for each scale. COSMIN procedure for the
124 recommendations of scales in systematic reviews [24] was used to guide the recommendations
125 made.

126 **Results**

127 **Study selection**

128 The initial search resulted in 9645 records and the re-executed search identified an additional 2339
129 records (Figure 1). Following removal of duplicates, there were 7811 records. A further 24 records
130 were identified through screening reference lists and manual searching. After screening of titles and
131 abstracts, 185 remained for full-text screening. Following full-text screening, fifty-seven publications
132 of 18 distinct scales (and 39 variations) were identified as meeting eligibility criteria (Supplementary
133 Table S1 in Additional File 4). Many publications reported multiple studies, even for the same
134 measurement property, for example, where the measurement property was assessed and reported
135 for different populations or different versions of the same scale, or where different methods were
136 used to assess the same measurement property.

137 The measurement properties and study characteristics are reported in Supplementary Table S2 in
138 Additional File 4. Seven apathy-specific scales were identified: The Apathy Evaluation Scale (AES)
139 [33]; Apathy Inventory (AI) [34]; Apathy Motivation Index (AMI) [35]; Starkstein Apathy Scale (AS)
140 [36]; Dementia Apathy Interview and Rating (DAIR) [37]; Dimensional Apathy Scale (DAS) [38]; The
141 Lille Apathy Rating Scale (LARS) [39]. Apathy subscales were present in eleven global scales designed
142 to assess a variety of constructs (such as dementia severity, and neuropsychiatric symptoms):
143 Alzheimer's Disease and Related Dementias mood scale (ADRD) [40,41]; Behavioural and Mood
144 Disturbance Scale (BMDS) [42]; Behavioral Syndromes Scale for Dementia (BSSD) [43]; Dysexecutive
145 Questionnaire (DEX) [44]; Frontal Systems Behavior Scale (FrSBe) [45]; Geriatric Depression Scale
146 (GDS) [46,47]; Behavioural Rating Scale for Geriatric Inpatients (GIP) [48]; Index of Mental Decline
147 (IMD) [49]; Key Behaviours Change Inventory (KBCI) [50]; Neuropsychiatric Inventory (NPI) [51];
148 Unified Parkinson's Disease Rating Scale (UPDRS) [52]. Only the measurement properties of apathy
149 subscales were assessed, not the overall global scale. Of the publications that met the inclusion
150 criteria, there was one each that pertained to the AD-RD, BMDS, BSSD, DEX, GIP, IMD, KBCI, AMI,

151 DAIR, two regarding the FrSBe and GDS, three regarding the AI and LARS, four regarding the UPDRS,
152 five regarding the DAS, eight regarding the AS, nine regarding the AES, and twelve regarding the NPI.
153 The majority of scales required respondents to select responses from a Likert scale, in relation to
154 various questions or items. Number of scored items in the scales ranged from one to 33. Recall
155 periods ranged from one week to one month, with some scales not specifying a recall period, or
156 specifying since the onset of a disease.

157 Risk of Bias

158 Results of the individual studies and their risk of bias ratings are reported in the online
159 supplementary material (Supplementary Tables S3 – S5 in Additional File 4). No studies assessed
160 cross-cultural validity, so this is not discussed nor included in the tables. Literature pertaining to
161 development was obtained for all scales except the DEX, GIP, and FrSBe. Few additional content
162 validity studies were available that met the eligibility criteria. Most content validity and development
163 studies had indeterminate results, due to a poorly reported or inadequate method. Twenty-seven
164 studies of structural validity across 16 publications met the inclusion criteria [33,37,53–66]. Three
165 studies had very good methodological quality, as most studies used exploratory factor analysis or
166 principle component analysis to assess structural validity, instead of the preferred confirmatory
167 factory analysis or item response theory methods. Internal consistency was assessed by 31
168 publications [33,34,36,37,43,53,55–79] and was considered a valid assessment (i.e. the scale was
169 based on a reflective model) in 38 studies. Some results were indeterminate due to lack of evidence
170 that the scale was unidimensional, and therefore uncertainty regarding whether internal consistency
171 should apply. There were 38 inter-rater or test-retest reliability studies
172 [33,34,36,37,40,42,43,53,65,70,73,76,77,79–88] from 23 publications. None were of very good
173 methodological quality, and just one was of adequate quality. Methodological quality of reliability
174 studies was mostly limited as a result of not using the optimal statistical method, such as the use of
175 Kappa rather than weighted Kappa, or Pearson or Spearman correlation instead of Intraclass
176 Correlation Coefficient (ICC). Where the most appropriate method was used, the model or formula of
177 ICC or weighted Kappa was often not reported. Six studies of measurement error were conducted
178 across four publications [37,56,77,82]. For all but one study, it was not possible to draw conclusions,
179 due to lack of appropriate anchor-based estimates of Minimal Important Change (MIC) for any of the
180 scales. One hundred and eighty studies of hypothesis testing for construct (including convergent,
181 divergent and known-group) validity that met the eligibility criteria were found from 45 publications
182 [33,34,36,37,43,49,51,53–58,60,62,65,67–71,73–75,77–81,83,85,87–100]. Most reported p values,

183 but not effect sizes, and 21 studies had indeterminate results due to not reporting sufficient
184 information.

185 Synthesis of results

186 A synthesis of the results of measurement properties per scale, including quality rating and GRADE
187 ratings for older adults and people with dementia is provided in table 1.

188 Apathy specific scales

189 *Apathy Evaluation Scale (AES)*

190 The AES is an 18-item apathy scale based on informant-report; self-report (AES-S) or clinician
191 assessment. Nine publications regarding the validity or reliability of the AES met the inclusion
192 criteria. The AES had sufficient content validity, though, like other studies, the evidence for this was
193 very low. There was moderate evidence for sufficient hypothesis testing for construct validity and
194 structural validity. The latter result limited the quality of evidence for sufficient internal consistency
195 (Cronbach's α = .86 to .95) to moderate also. There was low to moderate evidence for sufficient
196 reliability, except of the AES-S in people with dementia, where test-retest reliability was insufficient.
197 The only measurement property that the AES lacked evidence for was measurement error.

198 *Dimensional Apathy Scale (DAS)*

199 The DAS is a 24-item scale, made up of three subscales: executive, emotional and initiation. There is
200 a self-rated and proxy version, and a shorter proxy version (b-DAS), which retains the three subscales
201 across just nine items. Five articles investigating the DAS (including b-DAS) met the inclusion criteria
202 [62,63,74–76]. There was very low evidence of sufficient content validity of the DAS, including b-DAS,
203 and sufficient test-retest reliability, however this evidence came from a single study of the b-DAS so
204 conclusions may not be generalizable to the full version. There was moderate to high quality
205 evidence of sufficient hypothesis testing for construct validity. Structural validity and internal
206 consistency were not relevant due to this scale's formative nature, and there was no evidence for
207 measurement error.

208 *Lille Apathy Rating Scale (LARS)*

209 The LARS was developed to screen for and assess changes in apathy in people with Parkinson's
210 Disease, and was originally designed as a clinician-rated scale based on observations and answers
211 provided in an interview with the participant. Three articles of the LARS met the inclusion criteria
212 [65,70,83]. There was very low evidence of sufficient content validity, low to moderate evidence of

213 sufficient reliability, and high quality evidence for sufficient hypothesis testing for construct validity.
214 As with the DAS, structural validity and internal consistency were not relevant due to this scale's
215 formative nature, and there was no evidence for measurement error.

216 *Dementia Apathy Interview and Rating (DAIR)*

217 The DAIR was developed to assess apathy in people with dementia. One article met the inclusion
218 criteria [37]. There was very low evidence for inconsistent content validity of the DAIR in older adults,
219 and low evidence for inconsistent content validity in people with dementia. There was very low to
220 moderate evidence for sufficient structural validity, and low to moderate evidence of internal
221 consistency. There was very low evidence for sufficient test-retest reliability, and measurement
222 error, and low to high quality evidence of sufficient hypothesis testing for construct validity.

223 *Apathy Inventory (AI)*

224 The AI is a 3-domain apathy scale, initially created as a self or informant report via face-to-face
225 interview and developed for older adults and people with neurological disorders. Three articles of
226 the AI met the inclusion criteria [34,72,80]. Evidence for content validity and hypothesis testing for
227 construct validity was inconsistent. There was low evidence for sufficient reliability, and no
228 conclusive evidence for structural validity or internal consistency.

229 *Apathy Scale (AS)*

230 The AS is a 14-item apathy scale, administered through self-report or informant-report, via interview.
231 An 11-item paper and pencil version (AS-HC) without sub-questions has also been produced [58].
232 Eight articles of the AS met the inclusion criteria [36,58–61,77,90,91]. Despite the high quality
233 studies, the results regarding structural validity were inconsistent. The AS-HC however had moderate
234 to low evidence for sufficient structural validity and internal consistency. There was very low
235 evidence of sufficient content validity and reliability, and low to moderate quality evidence for
236 sufficient hypothesis testing for construct validity of the AS. There was no conclusive evidence for
237 internal consistency or measurement error.

238 Global scales with an apathy subscale

239 *Neuropsychiatric Inventory (NPI)*

240 The NPI is a well-known scale for assessing neuropsychiatric symptoms in people with dementia.
241 Each subscale of the NPI represents a different symptom, of which apathy is one. Originally designed
242 as a proxy assessment administered via interview, the NPI now has many variations, including those

243 which score the screening or sub-questions, as in the NPI-Alternate (NPI-A), and NPI-Clinician (NPI-C).
244 The NPI was the most studied scale, with 12 articles meeting the inclusion criteria [51,66,73,79,84–
245 88,94–96]. Content validity of the original NPI apathy subscale was inconsistent, as the emotional
246 domain was missing from the screening questions. In contrast, the NPI-C had sufficient content
247 validity. The NPI-A had sufficient structural validity with moderate to very low evidence, and there
248 was moderate evidence for sufficient internal consistency in people with dementia. There was very
249 low to low evidence of reliability for the original NPI. The NPI-C had better evidence of reliability,
250 with low and moderate evidence for sufficient interrater reliability. Hypothesis testing for construct
251 validity was found to be insufficient for the original NPI, supported by high quality evidence, and for
252 the NPI-C, evidence was inconsistent. The NPI-A lacked conclusive evidence for hypothesis testing for
253 construct validity, construct validity, and reliability, whilst the NPI-C and NPI had no conclusive
254 evidence for structural validity, internal consistency or measurement error of the apathy subscales.

255 *Behavioral Syndromes Scale for Dementia (BSSD)*

256 The BSSD is a measure of neuropsychiatric symptoms, which contains an apathy subscale consisting
257 of seven items, for which one publication met the inclusion criteria [43]. There was very low evidence
258 for sufficient content validity, and inconsistent reliability for face-to-face administration, with
259 insufficient reliability when administered by telephone. There was very low to moderate evidence of
260 sufficient hypothesis testing for construct validity, however, results should be interpreted with
261 caution, as no studies of convergent validity were included. There was no conclusive evidence for the
262 remaining measurement properties (structural validity, internal consistency, measurement error).

263 *Dysexecutive Questionnaire (DEX)*

264 The DEX was developed as part of the behavioural assessment of the dysexecutive syndrome test
265 battery. One publication of the DEX met the inclusion criteria [81]. There was inconsistent content
266 validity, very low evidence for sufficient test-retest reliability and moderate to high quality evidence
267 of inconsistent hypothesis testing for construct validity. There was no conclusive evidence for the
268 remaining measurement properties (structural validity, internal consistency, measurement error).

269 Scales with limited evidence

270 The AD-RD, IMD and UPDRS all had evidence regarding just one measurement property. The AD-RD
271 had very low evidence for sufficient test-retest reliability, the IMD had very low to low evidence of
272 sufficient hypothesis testing for construct validity, and the UPDRS had moderate evidence for
273 inconsistent hypothesis testing for construct validity.

274 There was low to very low evidence of insufficient hypothesis testing for construct validity of the AMI
275 and whilst there was low evidence for sufficient content validity, it is worth noting that some items
276 were too conflated with cognition or disinhibition (e.g. “I get things done when they need to be
277 done, without requiring reminders from others”).

278 The BMDS, FrsBe, GIP, KBCI and three-item subscale of the GDS (GDS-3a) all had inconsistent content
279 validity and evidence regarding one other measurement property, although for all cases evidence for
280 content validity came from researcher ratings only due to absent or indeterminate development and
281 content validity studies. Both the BMDS and GIP had very low evidence for sufficient reliability and
282 inconsistent content validity, with only 55% and 44% of items relevant to apathy respectively. Items
283 and response options of the BMDS created confusing double negatives, and the emotional domain of
284 apathy was missing from both the BMDS and the GIP. There was very low to low evidence of
285 sufficient hypothesis testing for construct validity, and in all three versions of the FrSBe rated by
286 reviewers, none had the required $\geq 85\%$ of items relevant to apathy, due to items related to personal
287 hygiene that could be conflated with other impairments. There was moderate to low evidence of
288 insufficient hypothesis testing for construct validity for the GDS-3a and its inconsistent content
289 validity was due to inclusion of items too conflated with physical ability, and lack of
290 comprehensiveness. Despite similar inclusion of items that could be conflated with physical ability
291 and dysphoria, the six item subscale of the GDS (GDS-6a) had sufficient content validity, as
292 comprehensiveness and comprehensibility were sufficient. The GDS-6a also had moderate to low
293 evidence of sufficient hypothesis testing for construct validity. The KBCI had low to very low evidence
294 of sufficient hypothesis testing for construct validity, and inconsistent content validity due to some
295 items not being sufficiently relevant to older adults and people with dementia (e.g. “has a lot of get-
296 up-and-go”), and others lacking clear comprehensibility (e.g. “is enterprising”). The results regarding
297 hypothesis testing for construct validity for the IMD, KBCI, GDS-6a and FrSBe should be interpreted
298 with caution, as no convergent validity studies met the criteria, and convergent validity is a superior
299 indicator of construct validity than divergent or known-group validity [27].

300 **Discussion**

301 According to COSMIN guidelines, scales should be recommended if they have sufficient content
302 validity, at least low-level evidence for sufficient internal consistency, and no high-quality evidence
303 for insufficient properties. The AES, AMI, AS, DAS, GDS-6a, LARS and NPI-C all had sufficient content
304 validity in older adults and people with dementia, but the AS, GDS-6a and NPI-C did not have
305 evidence for sufficient internal consistency. The AES had sufficient internal consistency, though the

306 AMI, DAS and LARS were based on a formative model, so internal consistency was not applicable.
307 Therefore, the AES was the only scale that met the COSMIN criteria for a recommended scale. The
308 (original) NPI was the only scale to meet COSMIN criteria for a scale that should not be
309 recommended for use due to high quality evidence for insufficient hypothesis testing for construct
310 validity. All other scales could potentially be recommended, depending on further research.
311 However, we argue that the BMDS and GIP are also inappropriate for assessing apathy in older adults
312 and people with cognitive impairment due to inclusion of too many items that are not relevant and
313 conflate apathy with cognition.

314 This review considered both apathy specific scales and apathy subscales derived from a global
315 assessment, as, though the latter may be designed for screening purposes, as in the NPI, they are
316 often used in place of full assessments, from which conclusions are drawn: for example, the NPI-
317 apathy subscale has been recommended as a primary outcome measure in clinical trials [101].
318 Therefore, it was deemed necessary to assess both types of measures to create a sufficiently
319 comprehensive review of the evidence for all apathy measures available that may be used to assess
320 apathy specifically in people with dementia and older adults. It is worth noting that the best quality
321 apathy measures were all apathy specific scales, rather than those derived from a global measure.
322 This highlights the importance of apathy specific measures, and may suggest that apathy subscales
323 derived from global instruments (such as the apathy subscales of the BMDS, BSSD, GDS, IMD, KBCI,
324 UPDRS, and NPI) should not be used to assess apathy in research or clinical practice, unless followed
325 by further assessment. However, there is not currently sufficient evidence to make these
326 conclusions, except for the NPI. The finding that the NPI should not be recommended for assessing
327 apathy contrasts with its popularity and previous recommendations [20,101]. Our study found that
328 the NPI apathy subscale had insufficient construct validity and inconsistent content validity,
329 suggesting it assesses something other than apathy, which expands previous studies which
330 concluded it had uncertain validity [21]. Whilst this could be due to the low quality of convergent
331 validity studies, (which were all of inadequate quality), divergent validity studies also supported this
332 finding, as they showed a high correlation with depression, suggesting the NPI apathy subscale may
333 conflate apathy with depression. It is important to note that the NPI was designed as a quick
334 assessment tool for numerous neuropsychiatric symptoms [51], and therefore it is perhaps not
335 surprising that it does not offer a comprehensive and targeted assessment of apathy specifically.
336 Therefore, the NPI may be best used as a screening tool, but not as an outcome measure or full
337 clinical assessment of apathy in older adults or people with cognitive impairment.

338 This systematic review applied a wider search strategy and eligibility criteria than previous systematic
339 reviews [21,22], resulting in the inclusion of a larger number of studies, allowing more evidence to
340 contribute to the results. Despite the numerous studies of measurement properties identified by this
341 review, evidence across all measurement properties was often of low or very low quality. In
342 particular, many development and content validity studies failed to report a systematic process of
343 how items were produced or refined, did not involve patients, carers, or members of the public, or
344 did not provide sufficient detail (e.g. even when it was clear that participants were involved in
345 assessing these properties, it was not clear what aspects [such as items, recall period, instructions,
346 response options] of the scale participants were consulted about). As such, the included publications
347 offered little evidence for content validity, with all but two studies result's being indeterminate, and
348 no study exceeding doubtful methodological quality. As a result, content validity was largely
349 determined entirely by reviewer ratings of the scale itself. COSMIN's reviewer rating technique
350 ensured a validity rating was possible, even in the presence of insufficient information from the
351 development and content validity studies. However, this also meant that content validity conclusions
352 were largely based on very low evidence.

353 Furthermore, COSMIN guidelines do not advise how to recommend studies of scales based on a
354 formative model, which discounted three scales (the AMI, DAS and LARS) from the
355 recommendations. As such, the COSMIN guided recommendations of measures is to be taken with
356 caution in this study. Regardless of this, no single scale had overwhelmingly superior measurement
357 properties. The AES, DAS, and LARS all had sufficient content validity, reliability, hypothesis testing
358 for construct validity, and structural validity and internal consistency where applicable, in people
359 with dementia and older adults. The LARS was the scale with the best evidence for good
360 measurement properties, and was the only scale to have high quality evidence for at least one
361 measurement property in both older adults and people with dementia. However, the LARS may have
362 less desirable measurement characteristics, as both the self and informant versions involve
363 interviewer ratings, as well as respondent reports, and was the largest scale found by the review,
364 with 33 items assessing apathy, so requires more resources and could be burdensome. The AES had
365 the second most consistent quality evidence across measurement properties, and may have
366 preferable measurement characteristics, as there are versions that do not require trained raters, and
367 have fewer items. This is consistent with the recommendation of the AES made by others [20,101].
368 The DAS is a promising scale, with good evidence for sufficient measurement properties, with the
369 exception of reliability. The DAS also has desirable measurement characteristics, as the pencil and
370 paper based scale does not require interviews, and a short version is available.

371 Previous systematic reviews of apathy measures used QUADAS, which was designed only to assess
372 studies of diagnostic accuracy, and applied these to studies of a variety of measurement properties.
373 COSMIN on the other hand provides guidance and criteria for assessing the quality of and evidence
374 for a variety of measurement properties. The high standards set by the COSMIN guidelines and
375 criteria were however sometimes unattainable. For example in a development study, a lack of
376 justification of recall period and response options can prevent the results of a development study
377 being rated as sufficient, yet these aspects represent a small part of the scale, and are rarely
378 provided by even the best quality studies. As COSMIN quality criteria are binary, it risks over
379 simplifying the complexities of the true measurement properties and research evidence. Weighted
380 criteria which place greater emphasis on the items may be preferable for content validity
381 assessment. An alternative for assessing quality of the remaining measurement properties is that
382 used by Radakovic and colleagues [21] which rated each result on a scale of four to six possible
383 scores depending on the measurement property being assessed. However, this does not appear to
384 have been developed in a systematic way, unlike COSMIN criteria that were created following a
385 Delphi procedure.

386 Bias in systematic reviews can be minimised by duplicating all rating activities, however, due to the
387 large number of studies found by this review, this was impractical in this study. The duplication of
388 review for a portion of the included studies did however help discussions around what these flaws
389 may be, limiting subjective decisions. Bias was further minimised by following COSMIN guidelines,
390 and creating additional criteria where required, informed by PPI when applicable, that could be
391 followed for all scales.

392 This review did not restrict the eligibility criteria to people with a diagnosis of dementia or restrict
393 the age criteria to all adults (instead, choosing that at least 50% should meet the criteria). This meant
394 that some studies included participants with various diagnoses, such as Parkinson's Disease and
395 depression, and included some participants that were younger than 65. Therefore, the results may
396 be less applicable to the population we set out to study. However, populations do not neatly
397 segment, and by opting for a more liberal inclusion criteria, we were able to include a wide variety of
398 studies that may not otherwise have been included. Furthermore, the GRADE approach to
399 determining the quality of evidence for each measure takes into account the directness of the
400 results, so evidence that was less direct (i.e. studies in other populations) was marked down
401 accordingly.

402 Apathy is a multidimensional phenomenon, including behavioural, cognitive, social or emotional
403 domains [1,102], and so it is expected that a comprehensive apathy scales should assess all these
404 aspects of apathy. For this reason, we did not include studies of scales that only assessed a part of
405 the apathy construct, such as studies that investigated the separate sub-scales of the LARS and DAS.
406 It is possible that the best assessment of apathy is through a combination of scales that assess
407 different individual apathy subdomains, which could be used alongside direct observational methods,
408 such as accelerometers and other experimental methods, that have recently been used to assess
409 certain aspects of apathy such reduced goal directed behaviour [103]. Future studies could consider
410 the evidence for assessing each individual domain of apathy separately.

411 Conclusion

412 A number of apathy scales of varying quality are available and have been validated in an older adult
413 and dementia community-dwelling population. The development of scales was generally poor, due to
414 lack of transparency and systematic approach in eliciting and refining items and developing the other
415 measurement aspects such as recall period and response options. Future development of scales
416 should include a clear and systematic approach at all stages, and involve patients or members of the
417 public as well as professionals to ensure good content validity. The NPI is not recommended for
418 apathy assessment, except as a screening tool. The LARS has good measurement properties, so is
419 recommended for use in use in older adults and people with dementia and MCI studies with
420 sufficient resources. The DAS, in particular the resource efficient b-DAS, is a promising scale that
421 requires more research into its properties, particularly reliability. The AES has good measurement
422 properties and characteristics and is recommended for use in older adults and people with dementia
423 and MCI especially in circumstances of limited resources and to limit responder burden.

424

425

426 **Statements**

427 **Acknowledgement**

428 The authors would like to thank the PPI members Marianne Dunlop, Maureen Godfrey, and Morag
429 Whitworth, as well as the wider 'dementia, frail older people and end of life care' PPI group at the
430 University of Nottingham, for their important insights. We would also like to thank the translators
431 Ester Bellavia, Marta Castro Rodriguez, and Kenichi Sakuda, and the senior research librarian Jane
432 Grogan for their contributions.

433 **Statement of Ethics**

434 As this was a systematic review of published studies, it was not necessary to gain ethical approval.

435 **Conflict of Interest Statement**

436 The authors have no conflicts of interest to declare.

437 **Funding Sources**

438 This work forms part of a PhD project, partially funded by School of Health Sciences and the Division
439 of Rehabilitation, Ageing and Wellbeing, University of Nottingham. This article also presents
440 independent research funded by the National Institute for Health Research (NIHR) under its
441 Programme Grants for Applied Research Programme (Reference Number RP PG 0614 20007). The
442 views expressed are those of the author(s) and not necessarily those of the NIHR or the Department
443 of Health and Social Care.

444 **Author Contributions**

445 CB conceived of the study with supervision from RH, SG and VvdW. CB developed and conducted the
446 search and pre-screening. CB, VvdW and CaB screened the studies for eligibility, and CB, SG and
447 VvdW reviewed the included studies for their risk of bias and quality. CB drafted the manuscript and
448 all authors edited the text and approved the final manuscript.

449

Figure Legends

Fig. 1. PRISMA diagram of study selection process

Tables

Table 1. Overall findings and GRADE

| Measure | Content validity | | Structural Validity | | Internal Consistency | | Reliability | | Measurement error | | Hypothesis testing for construct validity | |
|---------|---|--|---|------------------------|--|------------------------|------------------------|--------------------------|---------------------|------------------------|---|---|
| | Summary of findings | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings <i>Cronbach's α</i> | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings <i>N hypotheses confirmed / tested (%)</i> | Quality rating & GRADE |
| AD-RD | DS: Indeterminate | ? | | | | | r=.72 | + (Very Low; Very Low) | | | | |
| AES | DS: Indeterminate RR: Sufficient | + (Very low; Very low) | 1 main apathy factor with smaller factors of various description. | + (Moderate; Moderate) | .86 to .95 | + (Moderate; Moderate) | r/ICC= .72 to .94 | + (Moderate; Low) | SEM= 2.7 to 2.9 | ? | 48/69 (70%) | + [†] (Moderate ^Δ ; Moderate [†]) |
| AI | DS: Indeterminate RR: Inconsistent | +/- (Very low; Very low) | | | .83 to .96 | ? | Kappa/ICC = .96 to .99 | + (Low; Low) | | | 5/8 (67%) | +/- [†] (Moderate; High) |
| AMI | DS: Indeterminate RR: Sufficient | + (Very low; Very low) | | | .86 | * | | | | | 0/2 (0%) | - (Low; Very low) |
| AS | DS: Indeterminate CVS: Indeterminate RR: Sufficient | + (Very low; Very low) | 1 to 3 factors. | +/- (High; High) | .69 to .94 | ? | r/ICC=.78 to .90 | + (Very Low; Very Low) | SEM= 2.34 | ? | 8/12 (67%) | + ^{††} (Low ^Δ ; Moderate [†]) |
| BMDS | DS: Indeterminate RR: Mixed | +/- (Very low; Very low) | | | | | r=.90 | + (Very Low; Very Low) | | | | |
| BSSD | DS: Indeterminate RR: Sufficient | + (Very low; Very low) | | | | | ICC= .65 to .85 | +/- (Very Low; Very Low) | | | 1/2 (50%) | - [†] (Very low ^Δ ; Moderate [†]) |
| DAIR | DS: Inconsistent RR: Inconsistent (OA); Sufficient (PwD) | +/- (Very low; Low) | 1 factor | + (Very low; Moderate) | .89 | + (Low; Moderate) | r=.85 | + (Very Low; Very Low) | 100% agreement | + (1; 1) | 3/4 (75%) | + (Low; High) |
| DAS | DS: Indeterminate RR: Sufficient | + (Very low; Very low) | 3 factors: cognitive; behavioural; emotional. | * | .81 to .93 | * | ICC=.84 | + (Very Low; Very Low) | | | 10/13 (77%) | + (Moderate; High) |
| DEX | RR: Inconsistent | +/- (Very low; Very low) | | | | | ICC=.93 | + (Very Low; Very Low) | | | 2/4 (50%) | +/- (Moderate; High) |
| FrSBE | CVS: Indeterminate RR: Inconsistent | +/- (Very low; Very low) | 1 factor | ? | .80 to .88 | ? | | | | | 4/5 (80%) | + (Very low; Low) |
| GDS | GDS-3a: DS: Indeterminate RR: Inconsistent | GDS-3a: +/- (Very low; Very low) GDS-6a: + (Very low; Very low) | | | .51 | ? | | | | | GDS-3a: 0/2 (0%) GDS-6a: 3/3 (100%) | GDS-3a: - (Moderate; Low) GDS-6a: + (Moderate; Low) |

| Measure | Content validity | | Structural Validity | | Internal Consistency | | Reliability | | Measurement error | | Hypothesis testing for construct validity | | |
|---------|--|-----------------------------|---------------------|---------------------------|--|---------------------------------|---------------------------------|------------------------|---------------------------|------------------------|---|------------------------|-------------------------|
| | Summary of findings | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings <i>Cronbach's α</i> | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings | Quality rating & GRADE | Summary of findings <i>N hypotheses confirmed / tested (%)</i> | Quality rating & GRADE | |
| GIP | GDS-6a: DS: Indeterminate RR: Sufficient RR: Inconsistent | +/- (Very low; Very low) | | | | | | ICC= .72 to .83 | + (Very Low; Very Low) | MDD= 2.8 to 3.8 | ? | | |
| IMD | DS: Indeterminate | ? | | | | | | | | | | 3/3 (100%) | + (Very low; Low) |
| KBCI | DS: Inconsistent RR: Inconsistent | +/- (Very low; Very low) | | | | | | | | | | 6/7 (86%) | + (Low; Very low) |
| LARS | DS: Indeterminate RR: Sufficient | + (Very low; Very low) | 4 factors | * | .81 to .87 | * | r/ Kappa / ICC = .93 to 1.00 | + (Low; Moderate) | | | | 11/13 (85%) | + (High; High) |
| NPI | DS: Indeterminate RR: Inconsistent | +/- (Very low; Very low) | | | .82 to .83 | ? | r/ rs / ICC= .53 to .99 | + (Very Low; Low) | | | | 1/5 (20%) | - (High; High) |
| NPI-A | | | 1 factor | + (Very low; Moderate) | .91 | ? (OA); + (Moderate, PwD) | | | | | | | |
| NPI-C | CVS: Indeterminate RR: Sufficient | + (Very low; Very low) | | | | | | ICC= .87 | + (Low; Moderate) | | | 1/2 (50%) | +/- (Moderate; High) |
| UPDRS | DS: Indeterminate RR: Inconsistent | +/- (Very low; Very low) | | | | | | | | | | | |

Blank cells indicate no eligible studies or results.

Quality of measurement property: +, Sufficient; +/-, Inconsistent; -, Insufficient, ? Indeterminate.

GRADE: Quality of evidence rating in parentheses first indicates quality of evidence for older adults, then people with dementia.

* not applicable due to formative model.

† Greater emphasis placed on results of better quality (sub)studies

†† Greater emphasis placed on studies of convergent validity

^ Marked down for inconsistency

Abbreviations: AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; AES Apathy Evaluation Scale; AI, Apathy Inventory; AMI, Apathy Motivation Index; AS, Apathy Scale; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; CVS, Content Validity Study; DAIR, Dementia Apathy Interview Rating; DAS, Dimensional Apathy Scale; DEX, Dysexecutive Questionnaire; DS, Development Study; FrSBe, Frontal Systems Behavior Scale; GDS-3a, Geriatric Depression Scale 3 item apathy subscale; GDS-6a, Geriatric Depression Scale 6 item apathy subscale; GIP, Behavioral Rating Scale for Psychogeriatric Inpatients; IMD, Index of Mental Decline; KBCI, Key Behaviors Change Inventory; LARS, Lille Apathy Rating Scale; MDS-UPDRS, Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale; NPI,

Neuropsychiatric Inventory; NPI-A, Neuropsychiatric Inventory Alternative; NPI-C, Neuropsychiatric Inventory Clinician; OA, Older Adults; PwD, People with Dementia and MCI; RR, Reviewer Rating; UPDRS, Unified Parkinson's Disease Rating Scale

1

2 Appendices

3 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)"

4

5

6 Data Extraction Table

| Identifier | | | | | | Interpretability | | | | | |
|----------------------------|-----|--------|------|----------------|-----------------|--------------------|---------------------------|--|--|--|--|
| Number (for randomisation) | DOI | Author | Year | Title of study | Name of measure | % 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 | | | | | | | | | |
|-------------------------------|--------|-----------------|-----------------------------------|-------------------|----------------------|--|--|---|--|
| Information on response shift | Design | Sampling method | Setting (location, time, context) | Target population | eligibility criteria | N (in each sub-analysis where appropriate) | Measurement properties assessed (i.e. relevant COSMIN boxes to complete) | Further description of measure if needed (e.g. changes to original) | |
| | | | | | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |

| Participant Descriptives | | | | | | | | Content validity |
|---|---------------------|---------------------------------------|-----------|------------------------------|---|---|--|--|
| Country from which research was conducted | 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 <i>Describe method briefly, e.g. asking stakeholders (e.g. patients, carers, experts) about the relevance, comprehensiveness, comprehensibility of the measure</i> |
| | | | | | | | | |
| | | | | | | | | |

Psychometric properties

| Structural validity - (e.g. Comparative Fit Index or Tucker Lewis Index) | Internal consistency (e.g. Cronbach's alpha) | Reliability (e.g. intraclass correlation coefficient 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 |
|---|---|--|---|---|--|---|
| | | | | | | |
| | | | | | | |
| | | | | | | |

| | | | | |
|--|--|--|--|---|
| | | | | <p>Responsiveness (Longitudinal data only. Compared to another measure, compared across groups, or before & after intervention)</p> |
| | | | | |

7

8

9 COSMIN Risk of Bias

10

| 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 |

11 Adapted with permission from Mokkink et al.[27]

12 Each risk of bias checklist box is to be completed for each study that assesses that measurement
 13 property. Boxes 1 is to be completed for original development studies, whereas box 2 is to be
 14 completed for any additional content validity studies, or studies developing an established measure
 15 in a different population. Box 8 will not be completed for any study in this systematic review, as no
 16 gold standard measure of apathy exists. For details of how risk of bias is assessed for each
 17 measurement property, see Mokkink et al [27].

18 Full guidelines followed for this review are found in the comprehensive COSMIN user manual version
 19 1.0 dated February 2018 downloaded from: [https://www.cosmin.nl/wp-content/uploads/COSMIN-](https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018-1.pdf)
 20 [syst-review-for-PROMs-manual_version-1_feb-2018-1.pdf](https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018-1.pdf) and the content validity user manual
 21 version 1.0, downloaded from: [https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-](https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf)
 22 [for-content-validity-user-manual-v1.pdf](https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf)

23

24

25

26

27 Supplementary Tables

28 Table S.1. Overview of measures

| <u>Measure</u> | <u>N of publications meeting criteria†</u> | <u>Original intended...</u> a) <u>construct</u> b) <u>target population</u> c) <u>context</u> | <u>Version</u> | <u>Measurement characteristics (refers only to the apathy component of the scale)</u> | | | |
|----------------|--|---|--------------------------|--|--|------------------------|---|
| | | | | <u>Mode of administration & other administration information</u> | <u>Recall Period</u> | <u>Number of items</u> | <u>Scoring and Response options*</u> |
| AD-RD [40,41] | 1 [40] | a) Mood b) Moderate to severe AD c) Research or clinical | n/a | Interviewer-judgement, informed by observation and patient and carer interview | 7 days | 5 | Items rated for frequency on Likert scale (1 to 5, all options described) |
| AES [33] | 9 [33,53–57,71,78,89] | a) Apathy b) People with various clinical disorders or apathy, (with MMSE over 10 for patient reported version) c) Clinical | AES-C | Clinician-rated based on semi-structured interview with patient and observations. Bachelor level raters can conduct with 4-6 hours experience. 10 to 20 minutes to administer | 4 weeks | 18 | Items rated on Likert scale (1 to 4; all options described), and quantifiable items rated 1 to 4 based 0, 1-2, 2-3, 3 or more quantifiable instances. Requires verbal or nonverbal evidence of intensity. Total score is sum of item scores. Range 18 to 72. |
| | | | AES-I | Informant-report via paper and pencil 10 to 20 minutes to administer. | 4 weeks | 18 | Likert scale (1 to 4; all options described). Total score is sum of item scores. Range 18 to 72. |
| | | | AES-I (16 item versions) | Informant-report via paper and pencil | 4 weeks | 16 | Likert scale (1 to 4; all options described). Total score is sum of item scores. Range 18 to 64 |
| | | | AES-S | Self-report via interview (recommended) or paper and pencil 10 to 20 minutes to administer | 4 weeks | 18 | Likert scale (1 to 4; all options described). Total score is sum of item scores. Range 18 to 72. |
| | | | AES-12PD | Self-report | 4 weeks | 12 | Likert scale (1 to 4; all options described). Total score is sum of item scores. Range 18 to 48. |
| AI [34] | 3 [34,72,80] | a) Apathy b) Older adults with brain disorders c) Clinical | AI-C | Clinician opinion based on observations, and participant and informant answers to the AI when available. At least 20 minutes of observation | Since beginning of the disease, last clinical assessment, or other defined | 3 | Likert scale (0 to 4; 3 options described) Total score is the sum of item scores. Range 0 to 12 |

| Measure | N of publications meeting criteria† | Original intended... a) construct b) target population c) context | Version | Measurement characteristics (refers only to the apathy component of the scale) | | | |
|-----------|-------------------------------------|--|------------------------|---|---|-----------------|---|
| | | | | Mode of administration & other administration information | Recall Period | Number of items | Scoring and Response options* |
| | | | | | time period e.g. last four weeks. | | |
| | | | AI-I | Informant-report via interview | Since beginning of the disease or an otherwise specified time point | 3 | Screening questions: (Yes=0 or No) with follow-up questions rated on Likert scale (Frequency: 1 to 4; Severity: 1 to 3; all options described) Item score is Frequency x Severity. Range 0 to 12. Total score is the sum of items scores. Range 0 to 36. |
| | | | AI-S | Self-report via interview | Since beginning of the disease or an otherwise specified time point | 3 | Screening questions: 0=“Yes”; “No” with follow up question rated on a visual scale (1 to 12; end-points described). Total score is the sum of item scores. Range 0 to 36. |
| AMI [35] | 1 [67] | a) Apathy b) Healthy adults c) Research | n/a | Self-report via paper & pencil | 2 weeks | 18 | Likert Scale (0 to 4; all options described). Total score is sum of item scores. Range 0 to 72. |
| AS [36] | 8 [36,58–61,77,90,91] | a) Apathy b) Parkinson’s Disease c) Clinical | AS-HC | Self-report via paper and pencil | 4 weeks | 11 | Likert scale: (0 to 3; all options described). Total score is sum of item scores. Range 0 to 33 |
| | | | AS-I | Informant report via interview ~ 10 minutes to administer | 4 weeks | 14 | Likert Scale (0 to 3; all options described). Total score is sum of item scores. Range 0 to 42. |
| | | | AS-S | Self-report via interview | 4 weeks | 14 | Likert Scale (0 to 3; all options described). Total score is sum of item scores. Range 0 to 42. |
| | | | AS-S (13 item version) | Self-report via interview | 4 weeks | 13 | Likert scale: (0 to 3; all options described). Total score is sum of item scores. Range 0 to 39 |
| BMDS [42] | 1 [42] | a) Neuropsychiatric symptoms (behaviour & mood disturbances) b) Dementia c) Research | n/a | Informant report via interview | - | 11 | Likert scale (0 to 4; all options described) Total score is sum of item scores. Range 0 to 44. |
| BSSD [43] | 1 [43] | a) Neuropsychiatric symptoms (behavioural syndromes in AD) b) AD c) Clinical | n/a | Clinician-judgement based on information from interview with informant and informed by clinician observations | 1 week | 7 | Likert scale (0 to 6; all options described). Total score is not specified but presumable sum of item scores. |
| DAIR [37] | 1 [37] | a) Apathy b) Dementia (mild-moderate) c) Research and clinical | n/a | Interviewer-judgement based on informant reports. In person or over the phone. | 1 month | 16 | Main items rated on Likert scale by informant: (0 to 3; all options described) with follow-up questions to |

| Measure | N of publications meeting criteria† | Original intended... a) construct b) target population c) context | Version | Measurement characteristics (refers only to the apathy component of the scale) | | | |
|-------------|-------------------------------------|---|-----------|--|---------------|------------------------|---|
| | | | | Mode of administration & other administration information | Recall Period | Number of items | Scoring and Response options* |
| | | | | ~ 30 minutes administration time | | | determine if this was a change in apathy rated by the interviewer (no change; increase; decrease) Total score is sum of all items reflecting a change (more apathetic), divided by the number of items completed. |
| DAS [38] | 5 [62,63,74–76] | a) Apathy b) Neurodegenerative diseases specifically with motor disability c) Research and clinical | DAS-I | Informant reported via online or paper and pencil ~ 5 minutes to administer | 1 month | 24 (8 per subscale) | Likert scale (0 to 3; all options described). 'Executive', 'Initiation' and 'Emotional' subscales are scored by summing all items in sub-scale. Range 0 to 24. Total score is the sum of the subscale scores. Range 0 to 72. |
| | | | DAS-S | Self-reported via online or paper and pencil ~ 5 minutes to administer | 1 month | 24 (8 per subscale) | Likert scale (0 to 3; all options described). 'Executive', 'Initiation' and 'Emotional' subscales are scored by summing all items in sub-scale. Range 0 to 24. Total score is the sum of the subscale scores. Range 0 to 72. |
| | | | b-DAS | Informant reported via online or paper and pencil >5 minutes to administer | 1 month | 9 (3 per subscale) | Likert scale (0 to 3; all options described). 'Executive', 'Initiation' and 'Emotional' subscales are scored by summing all items in sub-scale. Range 0 to 27. Total score is the sum of the subscale scores. Range 0 to 27. (an awareness deficit rating is also present but not included in the total score) |
| DEX [44]^ | 1 [81] | - | - | - | - | - | - |
| FrSBe [45]^ | 2 [64,68] | - | FrSBe-14a | - | - | 14 | - |
| | | | FrSBe-11a | - | - | 11 | - |
| | | | FrSBe-6a | - | - | 6 | - |
| GDS [46,47] | 2 [69,92] | a) Depression b) Older adults c) Clinical screening | GDS-3a | Self-reported via paper and pencil (interviewer administered if required) | 1 week | 3 | Responses (Yes/No) that indicate depression are scored 1. Total score is sum of items. Range 0 to 3 |
| | | | GDS-6a | Self-reported via paper and pencil (interviewer administered if required) | 1 week | 6 | Responses (Yes/No) that indicate depression are scored 1. Total score is sum of items. Range 0 to 3 |

| Measure | N of publications meeting criteria† | Original intended... a) construct b) target population c) context | Version | Measurement characteristics (refers only to the apathy component of the scale) | | | |
|---------------|-------------------------------------|--|---------------------------------|--|---------------------------------------|----------------------------|--|
| | | | | Mode of administration & other administration information | Recall Period | Number of items | Scoring and Response options* |
| GIP [48]† | 1 [82] | - | GIP-subscale | - | - | - | - |
| | | | GIP-domain | - | - | - | - |
| | | | GIP-9a (subscale of the GIP-28) | Observation by health professional | 2 to 3 weeks | 9 | Likert scale (options not described) |
| IMD [49] | 1 [49] | a) 'Mental decline' or 'impairment' b) Older adults, particularly with dementia c) Research. (Possibly also for clinical evaluation of progression but should not be used for diagnosis) | n/a | Informant reported | Not reported | 3 | Items are rated using categories that are associated with weighted scores depending on the item. 0="Absent"; 2/3="Mild-moderate / discontinuous symptoms"; 4/5/6="Severe / continuous symptoms" Total score is sum of item scores. Range 0 to 15 |
| KBCI [50,104] | 1 [93] | a) Behaviour change b) Traumatic Brain Injury c) Clinical and research | KBCI-8a | Informant reported via paper and pencil | Not reported | 8 | Likert scale (all options described) Total score is the sum item scores but the scores attributed to the Likert scale and therefore the range is unspecified. |
| | | | KBCI-10a | Informant reported via paper and pencil | Not reported | 10 | Likert scale (all options described) Total score is the sum item scores but the scores attributed to the Likert scale and therefore the range is unspecified. |
| LARS [39] | 3 [65,70,83] | a) Apathy b) Parkinson's Disease c) Clinical and research? | LARS-C | Interviewer-judgement informed by patient self-report and interviewer observations during the interview with the patient | 4 weeks | 33 | Four items are based on 3 or 5 point Likert scales (all options described) For the remaining items, patient responses are categorised by the interviewer as 1 or -1 (all options described). Items are scored 0 if they are rated 'N/A' or the interviewer was not able to categorise the reply. Total score is the sum item scores. Range -36 to 36. |
| | | | LARS-I | Interviewer-judgement informed by informant-responses during the interview with the informant | 4 weeks | 33 | Five items are based on 3 or 5 point Likert scales (all options described) For the remaining items, informant responses are categorised by the interviewer as 1 or -1 (all options described). Items are scored 0 if they are rated 'N/A' or the interviewer was not able to categorise the reply. Total score is the sum item scores. Range -36 to 36. |
| NPI [51] | 12 [51,66,73,79,84-88,94-96] | a) Neuropsychiatric symptoms b) Dementia c) Research and clinical | NPI (original) | Informant rated via interview | 1 month (and represents a change from | 1 (but rated for frequency | Screening question (Yes=0; No), with follow-up questions using Likert scales, regarding severity (1 to 3; |

| Measure | N of publications meeting criteria† | Original intended... a) construct b) target population c) context | Version | Measurement characteristics (refers only to the apathy component of the scale) | | |
|-------------|-------------------------------------|--|-----------|--|---|---|
| | | | | Mode of administration & other administration information | Recall Period | Number of items |
| | | | | behaviour before the illness) | and severity) | all options described) and frequency (1 to 4; all options described). Total score is Frequency x Severity (a distress rating is also present but not included in total score) |
| | | | NPI-A | Informant rated via interview | 1 month (and represents a change from behaviour before the illness) | - Each item is rated for frequency on the same Likert scale as the original NPI. Total score is the sum of frequency scores. (Severity is also rated for the overall domain as per the original NPI procedure, but not included in the total score) |
| | | | NPI-C | Clinician-judgement, informed by information from the NPI with an informant and patient as well as other relevant information about the patient. Clinicians must have a minimum of two years' experience of NPSs in people with dementia | 4 weeks | 11 Each item is scored individually by informants, employing the Likert method as the original NPI, regarding frequency, severity and distress. Total score is the summation of frequency and severity item scores. A clinical rating method is also required: Each item is also rated by a clinician based on their clinical impressions, informed by the interview with the patient and informant, clinical notes and other carers, rated on Likert scale (0 to 3). Total score is the sum of these clinician rated item scores. Two separate total scores are obtained: one from the informant, one from the clinician. |
| UPDRS [52]^ | 4 [97–100] | - | UPDRS | - | - | 1 Likert scale (0 to 4; all options described). No total score calculation required as only 1 item present. |
| | | | MDS-UPDRS | Rater-judgement informed by interview with patient and / or informant | 1 week | 1 Likert scale (0 to 4; all options described). No total score calculation required as only 1 item present. |

29 † Number does not include development article where development article did not meet the inclusion criteria, even if it was later assessed for purposes of content validity

30 * Reverse coding is not included here

31 ^ Unable to obtain development article for rating

32 - Unable to obtain information

33 Abbreviations: AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; AES Apathy Evaluation Scale; AES-12PD, Apathy Evaluation Scale for Parkinson Disease; AES-C, Apathy
 34 Evaluation Scale Clinician; AES-I, Apathy Evaluation Scale Informant; AES-S, Apathy Evaluation Scale Self; AI, Apathy Inventory; AI-C, Apathy Inventory Clinician; AI-I, Apathy Inventory
 35 Informant; AI-S, Apathy Inventory Self; AMI, Apathy Motivation Index; AS, Apathy Scale; AS-HC, Apathy Scale Home Care; AS-I, Apathy Scale Informant; AS-S, Apathy Scale Self; b-DAS, brief-
 36 Dimensional Apathy Scale; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; DAIR, Dementia Apathy Interview Rating; DAS, Dimensional
 37 Apathy Scale; DAS-I, Dimensional Apathy Scale Informant; DAS-S, Dimensional Apathy Scale Self; DEX, Dysexecutive Questionnaire; FrSBe, Frontal Systems Behavior Scale; FrSBe-6a, Frontal
 38 Systems Behavior Scale 6-item apathy subscale; FrSBe-11a, Frontal Systems Behavior Scale 11-item apathy subscale; FrSBe-14a, Frontal Systems Behavior Scale 14-item apathy subscale;
 39 GDS, Geriatric Depression Scale apathy; GIP, Behavioral Rating Scale for Psychogeriatric Inpatients; IMD, Index of Mental Decline; KBCI, Key Behaviors Change Inventory; KBCI-8a, Key
 40 Behaviors Change Inventory 8 item apathy subscale; KBCI-10a, Key Behaviors Change Inventory 10 item apathy subscale; LARS, Lille Apathy Rating Scale; MDS-UPDRS, Movement Disorder
 41 Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale; NPI, Neuropsychiatric Inventory; NPI-A, Neuropsychiatric Inventory Alternative; NPI-C, Neuropsychiatric Inventory
 42 Clinician; UPDRS, Unified Parkinson's Disease Rating Scale

43

44 Table S.2. Overview of studies

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|---|----------|---|--------------------------------|---------------------------|--|---|
| [40] | AD-RD | English [^] | Reliability (test-retest). | "Approximately half lived in low-income housing." No confirmation from correspondence. | N=39 | nr | 79.33 (9.22 ; 55 to 96) | 49% | 17.21 (5.98, 3 to 24) | AD-RD apathy: 10.57 (3.88) |
| | | | Development (pilot study) | "Conducted in a dementia-specific day center and two skilled nursing facilities." No confirmation from correspondence | N=45 | Cognitive Impairment (type not specified) | 79.00 (8.37; 61 to 94) | 45% | 7.88 (6.47; 0 to 23) | nr |
| [40,41] | AD-RD | English [^] | Development (item elicitation via interviews) | Nursing home and day care. No confirmation from | N=39 | Carers of people with moderate to severe AD: Formal carers (N=19). | Nursing home: 85 (nr, nr) | 25% | nr | nr |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|---------------------|----------------------------|---|---|----------|---|--|---|--|--|
| | | | | correspondence regarding proportion. | | Informal carers (N=20). (Number of people with AD that were being interviewed about =20) | Day care: 81 (nr,nr) | | | |
| [53] | AES-C | Chinese | Structural validity; Internal consistency; Reliability (interrater & test-retest). Hypothesis testing (convergent, divergent & known groups). | Outpatients - confirmed all community dwelling via correspondence with author | N=92 | Major depressive disorder: Current Depression (CD; N=31) Remitted Depression (RD; N=30) Healthy Controls (Ctrl; N=31) | CD=66.13 (8.24) RD=67.83 (6.20); Ctrl=68.90 (6.20); | CD=45.16% RD=33.33% Ctrl=48.39% | nr | 2 means for each group reflect 2 different clinicians' ratings: CD=42.32 (10.45); 40.32 (11.92) RD=32.17 (8.27) ; 30.33 (7.46) Ctrl=27.87 (7.55); 28.55 (9.24) |
| [54] | AES-C; AES-I; AES-S | nr | Structural validity; Hypothesis Testing (convergent & divergent). | Community-dwelling (95.8%) and nursing home residents (4.2%). | N=121 | Dementia: AD (55.2%); MD (AD-DLB, 14.3%; AD-VaD, 5.7%); DLB (9.5%); VaD (5.7%), FtD (4.8%); 'other dementia' (4.8%). | 73.7 (9.4) | 47.1% | nr | nr |
| [55] | AES-C; AES-I; AES-S | English [^] | Structural validity; Internal consistency; Hypotheses testing (divergent & known groups). | Outpatient and community sample – confirmed all community dwelling via correspondence with author | N=75 | MCI (N=57); Cognitively normal (Ctrl N=18) | MCI: 74.5 (8.6, 53 to 86) Ctrl: 75.4 (6.0, 63 to 84) Total: 74.7 (8.0, 53 to 86) | MCI: 70.2% Ctrl: 22.2% Total: 58.7% | MCI: 27.3 (1.9, 23 to 30) Ctrl: 29.4 (0.8, 28 to 30) Total: 27.8 (1.9, 23 to 30) | AES-C: MCI: 60.9±7.7 (39–72) Ctrl: 68.4±4.3 (55–72) Total: 62.7±7.7 (39–72) AES-I: MCI: 61.1 (8.0, 42 to 72) Ctrl: 68.3 (4.5, 58 to 72) Total: 62.8 (7.9, 42 to 72) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|------------------------|----------------------------|---|---------------------------|---|--|--|--|---|---|
| | | | | | | | | | | AES-S: MCI: 63.3 (8.0, 40 to 72) Ctrl: 67.2 (4.2, 56 to 72) Total: 64.3 (7.4, 40 to 72) |
| [33] | AES-C; AES-I; AES-S | English [^] | Development (item elicitation and pilot); Structural validity; Internal consistency; Reliability (interrater & test-retest); Hypothesis testing (divergent & known groups). | Community-dwelling | N=123 (N=40 for pilot) (n/a for item elicitation) | Mixed sample: Healthy controls (Ctrl, N=31); Probable AD (N=21); Major Depression (Dep; N=30); Left Hemisphere Stroke (LHS, N=19); Right Hemisphere Stroke (RHS =22). | Ctrl: 68.3 (5.7,nr) AD: 70.8 (7.6,nr) Dep: 71.6 (5.7,nr) LHS: 66.2 (6.6,nr) RHS: 70.1 (5.0,nr) Total: 69.53 (6.03)* 55 to 85) | Ctrl: 45.16% AD: 47.62% Dep: 10.00% LHS: 57.89% RHS: 54.55% Total: 40.65% | Ctrl: 29.1 (1.1, nr) AD: 19.1 (6.5, nr) Dep: 28.0 (1.7, nr) LHS: 25.0 (4.6, nr) RHS: 26.9 (2.3, nr) | AES-C: reported separately for the 2 clinician ratings: Ctrl: 26 (6.2, nr); 25.8 (5.8, nr) AD: 44.4 (11.1, nr); 45.2 (11.7, nr); Dep: 40.5 (9.7, nr); 36.6 (8.3, nr) LHS: 31.9 (9.6, nr); 32.0 (11.7, nr) RHS: 34.7 (7.3, nr); 35.4 (9.6, nr) |
| | | | | | | | | | | AES-I: Ctrl: 26.3 (7.5, nr) AD: 49.1 (9.9, nr) Dep: 41.7 (15.0, nr) LHS: 28.1 (6.9, nr) RHS: 35.4 (10.9, nr) |
| | | | | | | | | | | AES-S: Ctrl: 28.1 (6.4, nr) AD: 35.5 (8.1, nr) Dep: 38.7 (9.8, nr) LHS: 32.2 (8.6, nr) RHS: 31.6 (6.7, nr) |
| [71] | AES-I; AES-I-16 | German | Internal consistency; Hypothesis Testing (divergent). | Community-dwelling. | N=100 (AES-I N=80.) | Dementia | 83.19 (8.32, 59 to 100, N=99) | 29% | 16.35 (7.60, 0 to 29, N=65) | AES-I: 31.74 (10.43, 8 to 48) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|---|--|--|--|--|--|---|
| [56] | AES-I; AES-S | Swedish | Structural validity; Internal consistency; Measurement error; Hypothesis Testing (divergent). | Outpatients and community sample – No confirmation from correspondence whether the outpatients were community-dwelling. | N=511 Complete AES-I N=367. Complete AES-S N=496. | Neurodegenerative disease and cognitive impairment: MCS (N=222. AES-I N=192. AES-S N=209) with subgroups of subjective cognitive decline (SCD, N=97) and MCI (N=125). Parkinson's Symptoms (PS, N=88. AES-I N=76. AES-S N=88), with subgroups of PD (PD, N=71); Parkinson's Disease Dementia or Dementia with Lewy Bodies (PDD-DLB, N=17). Ctrl (N=201. AES-I N=135; AES-S N=199) | MCS: 70 (6) MCI: 71 (6) PD: 67 (9) PDD-DLB: 74 (6) Ctrl: 75 (5) Total: 72 (7) | MCS: 44.3%* MCI: 52%* PD: 56.3%* PDD-DLB: 76.5%* Ctrl: 37.8%* Total: 46.4%* | <u>median (Q1 to Q3)</u> MCS: 29 (27 to 29) MCI: 27 (26 to 28) PD: 29 (27 to 30) PDD-DLB: 23 (20 to 24) Ctrl: 29 (28 to 30) Total: 29 (27 to 29) | AES-I MCS: 36.2 (10.6, nr) PS: 52.3 (11.4, nr) Ctrl: 28.7 (8.2, nr) Total: 36.6 (12.9, nr) AES-S MCS: 32.6 (8.8, nr) PS: 53.3 (10.6, nr) Ctrl: 28.0 (5.7, nr) Total: 34.2 (11.9, nr) |
| [89] | AES I; AES-S | Italian | Hypothesis Testing (divergent). | Outpatients – No confirmation from correspondence whether community-dwelling. | N=48 | Parkinson's Disease (PD) | 72.21 (9.01, nr) | 64.58%* | 22.83 (4.71, nr) | AES-I: 45.14 (13.09, nr) AES-S: 49.85 (10.37, nr) |
| [57] | AES-S | German [^] | Structural validity; Internal consistency; Hypothesis testing (convergent & divergent). | Author confirmed all community via correspondence. | N=665 | Parkinson's Disease Sub-sample of PD excluding comorbidities of dementia or depression (PDexclDd; N=339) | PD: 67.3 (7.90, nr) PDexclDd: 66.52 (7.96, nr) | PD: 67.9% PDexclDd: 66.52% | PD: 27.94 (2.23) PDexclDd: 28.47 (1.58) | PD: 30.63 (9.49) PDexclDd: 27.96 (7.59) |
| [78] | AES-12PD | German | Internal consistency; Hypothesis testing (convergent & divergent) | Data taken from a study that has been confirmed community via correspondence. | N=339 | Parkinson's Disease. (Sample split for analyses: Sample 1: N=170; Sample 2: N=169) Subsample of PDDd: N=42 | Sample 1: 68 (nr, nr) Sample 2: 68 (nr, nr) | Sample 1: 70.00% Sample 2: 70.41% | <u>median (Q1 to Q3)</u> Samples 1&2: 29 (nr, nr) | <u>median (Q1 to Q3)</u> AES: Samples 1&2: 27.0 (nr) AES-12PD: Sample 1: 17.0 (nr) Sample 2: 18.0 (nr) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|---|--|---|---|--|---|--|
| [34] | AI | French [^] | Development (item elicitation) Internal consistency; Reliability (test-retest & interrater), hypothesis testing (convergent, divergent & known groups) | n/a no participants. Author advised outpatients via correspondence not unable to confirm whether community dwelling. | n/a N=115. (Test-retest N=14). | n/a People with neurodegenerative disease or cognitive impairment: AD (N=60); PD without dementia (N=12), MCI (N=24) Ctrl (N=19). Test-retest: AD only. | n/a AD: 74.90 (7.11, nr) PD: 64.1 (11.9, nr) MCI: 71.67 (5.92, nr) Ctrl: 70.68 (8.21, nr) Total: 72.40 (7.52)* | n/a AD: 45.00 PD: 58.33 MCI: 29.17 Ctrl: 42.11 | n/a AD: 22.55 (3.98, nr) PD: 27.2 (3.5, nr) MCI: 28.21 (1.06, nr) Ctrl: 29 (nr, nr) | n/a <u>AI-I</u> AD: 9.20 (10.4, nr) PD: 8.00 (6.0, nr) MCI: 4.21 (8.6, nr) Ctrl: 1.05 (2.0, nr) <u>AI-S</u> AD: 3.74 (5.9, nr) PDexID: 9.10 (8.3, nr) MCI: 2.47 (3.8, nr) Ctrl: 1.51 (2.9, nr) |
| [80] | AI-C | Portuguese | Internal consistency; Reliability (interrater); Hypothesis testing (convergent). | nr, but confirmed all community via correspondence | N=175. | Mixed sample: AD (N=55) MCI (N=35) Dep (N=32) PD (N=30) Ctrl (N=23) | AD: 78.4 (nr, 61 to 95) MCI: 69.1 (nr, 60 to 86) Dep: 69.7 (nr, 55 to 88) PD: 66.5 (nr, 42 to 84); Ctrl: 67.3 (nr, 52 to 88) Total: 71.45* | Total: 34.3% | AD: 16.8 (nr, 0 to 27) PD: 26.9 (nr, 18 to 20) Dep: 24.3 (nr, 16 to 30) MCI: 25.4 (nr, 22 to 27) Ctrl: 29.1 (nr, 28 to 30) Total: 23.28* | AI scores nr. Apathy 'diagnosis' according to Robert et al criteria: AD: 63.6% PD: 20% Dep: 68.8% MCI: 0% Ctrl: 0% |
| [72] | AI-C | French | Internal consistency; Hypothesis testing (convergent). | Outpatients – No confirmation from correspondence whether community-dwelling. | N=40 | Cognitive Impairment AD (N=17); MCI (N=12); MD (N=8); VaD (N=2); DLB (N=1) | 77.5 (8.01, nr) | 45%* | 20 (6.73, nr) | nr |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|-------------------|----------------------------|---|--|----------------------------------|---|--|---|---|--|
| [67] | AMI | English | Internal consistency; Hypothesis testing (convergent) | Outpatients – No confirmation from correspondence whether community-dwelling. | N=149 | PD (N=102) Ctrl (N=147) | PD: 67.7 (8.1,nr) Ctrl: 66.1 (8.5, nr) All at least 18 to 80 | PD: 77.5% Ctrl: 70.75% | ACE-III: PD: 89.4 (9.0, nr) All at least over 50 Ctrl: nr | PD: 35.29% apathetic in at least one AMI subscale |
| [90] | AS-I | Portuguese | Content validity | Outpatients – No confirmation from correspondence whether community-dwelling. | N=11 | Dementia: AD (N=8); FtD (N=3); | AD: 78.3 (4.7) FtD: 55 (8.7) Total: 71.95 (5.59)* | AD: 50.00%* FtD: 33.33%* Total:45.45%* | nr for this sample. Total: 20.64 (3.85)* | 22.8 (8.4, 12 to 39) |
| | | | Hypothesis testing (convergent & divergent) | Population random sample – No confirmation from correspondence whether community-dwelling. | N=20 | Probable or Possible AD | 84.1 (5.8) | 30% | 17.4 (SD=4.7) | 23.6 (10.6; 9 to 40) |
| [59] | AS-S (14/13 item) | English [^] | Structural validity; Internal consistency. | nr, but confirmed all community via correspondence | N=226 | Parkinson's Disease, without dementia. | 65.02 (8.84, nr) | 66.70% | (N=7) 29.14 (0.69, nr) | 10.99 (6.26, nr) |
| [36] | AS-S | English [^] | Development | n/a, no participants | n/a | n/a | n/a | n/a | n/a | n/a |
| | | | Internal consistency; Reliability (interrater & test-retest); Hypothesis testing (known groups) | nr. Author unable to access the information. | N=50 (Reliability studies: N=11) | Parkinson's disease, grouped into sub-samples based on apathy and depression scores: PD, no apathy, no depression (PD; N=16) PD, with apathy, no depression, (PDa; N=6) PD, no apathy, with depression, (PDd; N=13) | PD: 67 (9, nr) PDa: 69 (7, nr) PDd: 62 (12, nr) PDa&d: 69 (8, nr) Total: 66.54 (9.26)* | PD: 50% PDa: 66% PDd: 57% Pa&d: 73% Total: 62%* | PD: 28.7 (1.1, nr) PDa: 28.3 (1.2, nr) PDd: 26.3 (4.6, nr) PDa&d: 25.4 (4.5, nr) Total: 27.04 (3.06)* | PD: 7.3 (2.8, nr) PDa: 17.1 (4.0, nr) PDd: 10.0 (2.0, nr) PDa&d: 19.5 (3.3, nr) Total =12.84 (2.87)* |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|---|------------------------------|---|--|---------------------------|---|---|
| | | | | | | PD, with depression and apathy (PDA&d; N=15) | | | | |
| [58] | AS-S AS-HC | Japanese | Structural validity; Internal consistency; Hypothesis testing (divergent). | “Home-care” recipients. Assumed community-dwelling | N=122 | Parkinson's Disease | 70.9 (7.8, nr) | 49.2% | nr | AS-S: 26.6 (8.12, nr) AS-S-11: 21.3 (6.88, nr) |
| [60] | AS-S | Norwegian | Structural validity; Internal consistency; Hypothesis testing (divergent). | nr. No confirmation from correspondence whether community-dwelling. | N=194 | Parkinson's Disease | 67.9 (9.0, nr) | 59.3% | 27.8 (2.3, nr) | 15.5 (4.6, 4 to 29) (median =15.0). |
| [77] | AS - S | Spanish [^] | Internal consistency; Reliability (test-retest); Measurement error; Hypothesis testing (divergent & known-groups) | Outpatients – No confirmation from correspondence whether community-dwelling. | N=211 (test-retest: N=71) | Parkinson's Disease | 67.5 (10.2, nr) | 65.5%* | Short Portable Mental Status Questionnaire of Pfeiffer: 1.3 (1.6, nr). | 12.7 (7.1, nr) |
| [61] | AS-S | English [^] | Structural validity; Internal consistency. | Outpatients. Confirmed community-dwelling via correspondence | N=233 | Parkinson's Disease and healthy controls PD (N=157) Ctrl (N=76) | PD: 67.64 (8.27, nr) Ctrl: 66.95 (8.73, nr) | PD: 68.15%* Ctrl: 44.74%* | Mattis dementia rating scale: PD: 138.48 (3.88,nr) Ctrl: 140.46 (3.24,nr) | PD: 11.59 (5.36,nr) HC: 9.21 (4.67,nr) |
| [91] | AS-S | Spanish | Internal consistency; Hypothesis testing (convergent; divergent) | nr. Unknown to corresponding author as data not collected. | N=60 | Advanced Parkinson's Disease | 68.02 (7.43; 50 to 81) | 60.70% | nr | 11.55 (6.49, 1 to 24) |
| [42] | BMDS | English [^] | Development (item elicitation) | n/a, no participants | n/a | n/a | n/a | n/a | n/a | n/a |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|--|---|---|--------------------------------|---------------------------|--|--|
| | | | Reliability (test-retest). | nr, but scale designed to assess people living in the community | N=38 (test-retest reliability N=18) | Dementia | 76 (nr, 59 to 87) | 23.68% | nr | 24.95 (9.30, nr) |
| [43] | BSSD | English | Development (item elicitation and pilot) | Item elicitation: n/a no participants Pilot: nr | nr | nr | nr | nr | nr | nr |
| | | | Internal consistency; Reliability (interrater & test-retest); Hypothesis Testing (divergent & known groups) | Outpatients – No confirmation from correspondence whether community-dwelling. | N=106 (hypothesis testing: N=83 to 97; reliability: N=20 to 21) | Alzheimer's Disease | 72.1 (9.8, 45 to 93) | 35% male | Modified MMSE: 26.2 (13.8, 0 to 52) | Global apathy / indifference =31.1% absent; 50.0% minimal to mild; 18.8% moderate to severe. raw scores nr. |
| [37] | DAIR | English [^] | Development (item elicitation and pilot); | nr | nr | Mixed sample: People with AD, their carers and clinical researchers. | nr | nr | nr | nr |
| | | | Structural validity; Internal consistency; Hypothesis testing (convergent & divergent) | nr Designed to assess people living in environments whose daily activities are not structured, suggesting community-dwelling. No confirmation from correspondence whether community-dwelling. | N=100 | Alzheimer's Disease | 75.00 (8.48; 52 to 92) | 50% | 18.55 (7.20; 3 to 29) (Unobtainable for 16%) | 1.19 (0.69, 0 to 3) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|--|---|---------------|--|--|--------------------------------------|--|--|
| [38] | DAS | English (assumed) | Development study (item elicitation) | n/a no participants involved in item elicitation | n/a | n/a | n/a | n/a | n/a | n/a |
| [75] | DAS | English^ | Internal consistency; Hypothesis testing (convergent; divergent) | Outpatients - all confirmed community via correspondence. | DAS-S N=68 | Parkinson's Disease without dementia and healthy controls: PD (N=34) Ctrl (N=34) | PD: 68.2 (9.2, nr) Ctrl: 66.1 (9.2, nr) | 44.12% | nr | PD: 25.8 (8.7, nr) Ctrl: 21.2 (7.0, nr) |
| | | | | | DAS-I N=60 | (sub-sample of those above) PD (N=30) Ctrl (N=30) | nr for this sub-sample | nr for this sub-sample | nr | PD: 25.1 (12.8, nr) Ctrl: 19.7 (9.5, nr) |
| [74] | DAS | English^ | Internal Consistency; Hypothesis testing (convergent & divergent) | Community-dwelling | N=157* | DAS-I Alzheimer's Disease and controls AD (N=102) Ctrl (N=55) | AD: 78.2 (8.5, nr) 82.4% aged 65 and over. Ctrl: 75.0 (6.1, nr) | AD: 51.0%* Ctrl: 50.9%* | AD (N=80): 22.0 (5.3, nr) Ctrl: nr | nr, but AES: AD: 51.7 (11.5, nr) Ctrl: 28.8 (5.2, nr) |
| | | | | | | DAS-S AD (N=55, sub-sample of those above) Ctrl (same as above, n=55) | AD: 77.5 (7.9, nr) Ctrl: 75.0 (6.1, nr) | AD: 50.9%* Ctrl: 50.9%* | nr | nr, but AES: AD: 38.9 (9.0, nr) |
| [62] | DAS-S | Italian | Structural validity, Internal consistency, Hypothesis testing (convergent, divergent & known groups) | Outpatients - all confirmed community via correspondence. | N=207 | Parkinson's Disease and controls PD (N=107) Ctrl (N=100) | PD: 66.02 (9.01, nr) Ctrl: 64.52 (8.79, nr) | PD: 60.75%* | PD: 27.63 (2.09, nr) | PD: 25.25 (12.76, nr) (Median (skewness)=23 (1.254)) Ctrl: 21.29 (8.35, nr) |
| [63] | bDAS | English | Structural validity | AD: Community-dwelling ALS: nr | N=204 | Neurodegenerative Disease AD (N=102) ALS (N=102) | AD: 78.2 (8.5, nr) ALS: 63.8 (11.0, nr) | AD: 51%* ALS: 70%* Total: 60%* | AD: (N=80): 22.0 (5.3, nr) ALS: nr Total: nr | nr for bDAS AES: AD: 51.7 (11.5, nr) ALS: 33.2 (10.8, nr) Total: 42.4 (14.4, nr) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|--|--|--------------------------------|---|--------------------------------|---------------------------|--|--|
| | | | | | | | Total: 71.0 (12.1, nr) | | | |
| [76] | bDAS | English [^] | Internal consistency; Reliability (test-retest). | All confirmed community via correspondence. | N=53 (reliability N=43) | ALS | 68.0 (7.5, nr) | 83.01%* | ECAS cognitive score: 107.0 (14.1,nr) | nr for total score DAS-I subscales: Executive: 6.1 (4.8, nr) Emotional: 8.9 (4.2, nr) Initiation: 12.1 (5.5, nr) b-DAS Executive: 2.0 (2.0, nr) Emotional: 2.9 (1.9, nr) Initiation: 4.3 (2.6, nr) |
| [81] | DEX | Japanese | Reliability (test-retest); Hypothesis testing (convergent & divergent) | Outpatients. | N=122 (reliability N=44) | Alzheimer's Disease | 72.0 (7.7, nr) | 37.70%* | 20.8 (2.0, nr) | nr |
| [68] | FrSBe-I | English [^] | Content validity (cognitive interview) | Outpatients - all confirmed community via correspondence. | N=10 | People attending neuropsychological evaluation. 90% had memory complaints. Diagnoses nr. | nr | nr | nr | nr |
| | | | Structural validity; Internal consistency; hypothesis testing (groups & divergent); | Outpatients - all confirmed community via correspondence. | N=494 | Mixed sample: Dementia: AD (19.3%*), VaD (4.9%); Dementia not otherwise specified (4.1%); MD (4.5%); FTD (4.1%); DLB (1.8%). PD (16.6%). MCI (12.5%). Cognitive disorder not otherwise specified (CDNOS, 8.8%). | 69.92 (13.96, 19 to 95) | 47.04%* | nr | Original FrSBe-apathy: PD=33.29 (12.71); AD =37.24 (10.18); Frontal impairment =38.18 (10.35) Revised FrSBe-apathy: PD=27.24 (10.13); AD =29.71 (7.83); Frontal impairment =30.21 (8.08) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|--|-------------------|--|--|---------------------------|--|--|
| | | | | | | Frontal stroke (7.2%). Head injury (2.1%). Other neurological disorder (<1%). | | | | |
| [64] | FrSBe-I | English [^] | Structural validity; Internal consistency | Outpatients - all confirmed community via correspondence. | N=304 | Older adults with memory complaints: Dementia (N=166) MCI (N=63) No definitive diagnosis (NDD; N=28) Ctrl (N=47) | 79.12 (8.05; 52 to 99) | 28.29%* | nr | 86.12 (24.39) |
| [46,47] | GDS-30 | English [^] | Development (Item elicitation and pilot study) | Item elicitation: n/a no participants Pilot: Community dwellers (N=20) and inpatients (N=51). | N=71 | Healthy older adults (Ctrl: N=20) Depressed older pts (Dep: N=51) | nr. All over 55. | nr | nr | nr |
| [92] | GDS-3A | Dutch [^] | Hypothesis testing (convergent validity) | Community-dwelling | Study 1 N=427 | Older adults with mild cognitive deficits | 81.3 (4.6, nr) All at least 75 and over | 39.8%* | <u>median (Q1 to Q3)</u> 26 (25 to 27) | GDS-3a score: 0 =52.8%; 1=30.7%; 2=12.2%; 3=4.4% AS: 11.3 (4.7) |
| | | | | | Study 2 N=1118 | Older adults with depressive symptoms | 81.8 (4.9, nr) All at least 75 and over | 38.9%* | <u>median (Q1 to Q3)</u> 28 (27 to 29) | GDS-3a: 0 =64.2%; 1 =25.6%; 2 =9.3%; 3 =0.89% AS: 7.5 (4.6, nr) |
| [69] | GDS-6A | English [^] | Internal consistency, Hypothesis testing (divergent & known groups) | Community-dwelling | N=140 | Mixed sample: Dementia: AD (29.3%); VaD (29.3%); MD (13.6%) Cognitive disorder not specified or MCI (CNS-MCI, 17.1%) Other (6.4%); None (2.1%) (2.2% nr) | 78.2 (7.23, nr) All at least 65 or over | 35.0%* | 24.86 (3.35, nr) | GDS-6a: 1.66 (1.39, nr) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|--------------------|----------------------------|--|--|--|--|--|---------------------------|--|--|
| [82] | GIP-a-s GIP-a-d | Dutch | Reliability (test-retest); Measurement error. | All confirmed community via correspondence. | N=109 Complete and analysed: N=56. | Mixed sample: Dementia: AD (82%); VaD (13%); Other dementia (3%); Other (affective disorder or other cognitive disorder, 2%) | <i>median (Q1 to Q3, range)</i> 80 (75.5 to 84, 53 to 96) | 42.2%* | <i>median (Q1 to Q3, range)</i> Cognitive Screening test: 13.3 (10.4 to 16, 3.5 to 20) Amsterdam Dementia Screening test 3: 0 (-2 to 1, -5 to 4) Amsterdam Dementia Screening test 5: 1 (-1 to 3, -5 to 8). | N=56: GIP-a-s: 2.2 (2.3, 0 to 9) GIP-a-d: 2.8 (3.5, 0 to 15) |
| [49] | IMD | Italian [^] | Development (item elicitation) Hypothesis testing (divergent) | n/a no participants | n/a | n/a | n/a | n/a | n/a | n/a |
| | | | | Sample 1: Some Community-dwelling and some institutionalised. Author unable to confirm proportion. | N=236 | nr, but at least some healthy older adults. Mild to moderate functional impairment (52.5%). Severe functional impairment (24.8%). | 74.2 (6.8, nr) | 40.6%* | 19.4 (4.3, nr) | nr |
| | | | | Sample 2: nr. Author unable to confirm. | N=203 | Dementia | 74.1 (5.56; 63 to 83) | 33.99%* | 19.7 (2.61, 15 to 23) | 5.4 (3.15) |
| [50,104] | KBCI | English [^] | Development (item elicitation) | nr | nr | People with TBI, their carers, and TBI rehabilitation specialists. | nr | nr | nr | nr |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|--|--|---|--|--|----------------------------------|--|---|
| | | | Development (item refinement) | panel1: nr. panel 2 & 3: n/a. | N=14 | Panel 1: carers for people with TBI (N=4) Panel 2: clinical psychologists (N=3) Panel 3: clinical neuropsychologists (N=7) | nr | nr | nr | nr |
| [93] | KBCI-a | English [^] | Hypothesis testing (divergent) | Outpatients. No reply from author. | N=97 | Mixed sample: Ctrl (31%) MCI (18%) Probable AD (7%) Other (depression, CDNOS, PD, DLB, and possible AD) | 72.34 (9.05, nr) | nr | 26.89 (2.63, nr) | nr |
| [39] | LARS | French; English | Development | n/a – no participants involved. | n/a | n/a | n/a | n/a | n/a | n/a |
| [83] | LARS - C | Spanish | Reliability (interrater & test-retest); Hypothesis Testing (convergent) | Community-dwelling (“non-institutionalised”) | N=151 (test-retest N=16, interrater N=21) | Dementia (Dem, N=101) and healthy controls AD (N=43) FtD (N=41) Primary Progressive Aphasia (N=17) Ctrl (N=50) | Dem: 74.3 (7.7, nr) Ctrl: 72.0 (9.7, nr) | Dem: 45.5%* Ctrl: 38%* | Dem: 21.59 (6.21, nr) Ctrl: 28.72 (1.42, nr) | Dem: -0.16 (18.50, nr) Ctrl: -29.54 (5.44, nr) |
| [70] | LARS-I | French [^] | Internal consistency; Reliability (interrater & test-retest); Hypothesis Testing (convergent) | Correspondence with author confirmed all community | N=60 (interrater N=34, test-retest N=29) | Parkinson’s Disease: PD without dementia (PDexclD, N=43) PD with dementia (PDD, N=17) | PDexclD: 64.74 (9.29, nr) PDD: 69.53 (9.06, nr) Total: 66.10 (9.23)* | PDexclD: 67.44%* PDD: 35.29%* | nr | -16.18 (11.99, nr) |
| [65] | LARS - C | Spanish | Content validity; Structural validity; Internal consistency; Reliability (interrater & test-retest); | nr. No confirmation from correspondence whether | N=200 (content validity and reliability N=30) | Parkinson’s Disease and healthy controls PD (N=130) Ctrl (N=70) | PD: 71.6 (8.1, nr) Ctrl: 69.4 (8.7, nr) | PD: 60.0%* Ctrl: 55.7%* | MEC: PD: 30.7 (3.8, nr) Ctrl: 33.3 (1.7, nr) | PD: -14.5 (9.1, nr) Ctrl: -25.0 (5.5, nr) |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|--|--|--|---|---|---|--|--|
| | | | Hypothesis testing (convergent & divergent) | community-dwelling. | | | | | | |
| [94] | NPI | Korean | Hypothesis Testing (known groups). | Assessment setting suggests outpatients. No confirmation from correspondence whether community-dwelling. | N=141 (test-retest N=29) | Dementia (N=92) and healthy controls: AD (N=43) VaD (N=32) FtD (N=11) Other dementia (N=6) Ctrl (N=49) | Dem: 67.5 (9.7, 38 to 85) Ctrl: 66.9 (8.4, 51 to 82) | Dem: 47.8%* Ctrl: 34.7%* | Dem: 17.5 (6.8, 0 to 29) Ctrl: 26.3 (2.3, 19 to 30) | NPI-apathy total nr. <u>Dem:</u> Prevalence: 77.2%. Frequency: 2.52 (1.67; 0 to 4) Severity: 1.75 (1.18; 0 to 3) <u>Ctrl:</u> Prevalence =6.1%. Frequency =0.06 (0.24; 0 to 1) Severity =0.06 (0.24; 0 to 1) |
| [51] | NPI | English [^] | Development (item elicitation and Delphi study of comprehensiveness) Reliability (interrater & test-retest) | Item elicitation: n/a no participants Delphi study: n/a professionals Community-dwelling | N=10 N=80 (interrater N=45, test-retest N=20) | Geriatric psychiatrists, behavioural neurologists, and neuropsychologists Dementia (Dem) and healthy controls: AD (N=20) VaD (N=9) Other dementia (N=11) Ctrl (N=40) | n/a 75.7 (56 to 90) | n/a Dem: 55.00%* Control: 50.00%* | n/a Dem: 19.2 (0 to 29) Control: 28.4 (25 to 30) | n/a NPI-apathy total nr. Frequency: 2.83 (1.55; 0 to 4) Severity: 1.35 (0.83; 0 to 3) |
| [85] | NPI | Icelandic | Reliability (test-retest); Hypothesis testing (known groups). | Community-dwelling | N=38 (test-retest N=6) | Dementia: AD (N=19) VaD (N=19) | 78.84 (6.66; 59 to 89) | 47% | 19.26 (5.95; 1 to 29) | nr for total sample. Reported separately for two different severity groups (N in each group nr). |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|---|--|---|--|--|-------------------------------|---|---|
| | | | | | | | | | | Less severe dementia: 4.69 (3.72, nr) More severe dementia: 7.45 (4.45, nr) |
| [73] | NPI | Farsi | Internal consistency; Reliability (interrater & test-retest); Hypothesis testing (convergent, divergent & known groups) | 51% living with family, suggesting at least majority community dwellers. No confirmation from correspondence | N=100. (interrater N=50, test-retest reliability N=30, hypothesis testing N=50) | Dementia and healthy controls. Dem (N=100) Ctrl (N=49) | Dem: 74.5 (8.3, 60 to 90) Ctrl: 74.3yrs (8.5) | Dem: 47% Ctrl: 51% | nr for total sample. Hypothesis testing (N=50): Dem: 11.3 (7.5, nr) Ctrl: 29.4 (1.0, nr) | NPI-apathy total nr. Prevalence: 74% Frequency 2.5 (1.7, nr) Severity 1.6 (1.1, nr) |
| [79] | NPI | Spanish | Internal consistency; Reliability (interrater); Hypothesis testing (convergent) | Outpatients – No confirmation from correspondence whether community-dwelling. | Total N=63. (interrater N=39) | Mixed sample: Dem (N=44) Dep (N=6) Ctrl (N=13) | 72.76 (9.67; 35 to 85) | 49.21%* | nr | NPI-apathy total nr. Prevalence: 56% |
| [95] | NPI | Greek | Hypothesis testing (convergent) | Outpatients. Author correspondence confirmed all community. | N=29 | Dementia | 71.05 (5; 60 to 84) | 60% | 12.4 (6.0; 0 to 24) | 5.8 (4.4, nr) |
| [86] | NPI | Chinese | reliability | Community dwelling | N=91 | Dementia and healthy controls. Dementia (Dem, N=62*): AD (N=41), VaD (N=16), Other (N=5) Ctrl (N=29) | Dem: 76.4 (7.0; 54 to 88). Ctrl: 74.9 (4.7; 68 to 86) | Dem: 22.58%* Ctrl: 72.41%* | Dem: 12.7 (5.9; 0 to 25.) Ctrl: 27.5 (2.2; 23 to 30.) | nr |
| [84] | NPI | Brazilian Portuguese | Reliability (interrater & test-retest) | Outpatients. Author correspondence confirmed all community | N=36 | Alzheimer's Disease | 78.78 (7.48) | 22%* | 7.06 (6.92) | NPI-apathy total nr. Severity: 5.31 (4.91) Frequency: 1 =33%, 2 =3%, 3 =64%. |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|---|--|--|---|---|--------------------------------|---------------------------|--|--|
| [96] | NPI | Dutch | divergent validity | 83.33% community-dwelling | N=24 | Mixed sample: Dementia: AD (N=19), FtD (N=1), MD (N=1) Stroke (N=2) Amnesic disorder (N=1) | 74.3 (10.4, nr) | 33.33%* | 21.5 (4.6; 12 to 29)." | nr |
| [66] | NPI-A | English [^] | Structural Validity; Internal consistency. | Outpatients. Author was unable to confirm whether community-dwelling. | N=124 | Dementia: AD (N=62) VaD (N=43) MD of AD+VaD (N=19) | 79.8 (6.1; 61 to 91) | 21.77%* | 22.6 (3.5; 13 to 29) | 8.89 (8.5, nr) |
| [87] | NPI-C | English [^] , French [^] , Greek [^] , Italian [^] , Hungarian [^] , Portuguese [^] , Spanish [^] | Content validity (further item elicitation and Delphi study) Reliability (interrater); Hypothesis Testing (convergent) | Item elicitation: n/a no participants Delphi study: n/a professionals | Delphi study: N=8 | Experts in dementia research | n/a | n/a | n/a | n/a |
| [88] | NPI-C | Portuguese | Reliability (interrater); Hypothesis Testing (convergent) | Author confirmed all community via correspondence | N=156 | Dementia | 76.7 (nr, nr) | 26.28%* | 17.2 (nr, nr) | NPI-C-aphathy total nr. AES (N=113): 33.1 (11.3; 0 to 51) |
| [52] | UPDRS | English | Development (item elicitation and review of comprehensibility) | n/a no participants involved | n/a | n/a | n/a | n/a | n/a | n/a |
| [100] | UPDRS | Spanish [^] | Hypothesis Testing (convergent) | Outpatients – No confirmation from correspondence whether community-dwelling. | N=168 (convergent validity N=164) | Parkinson's Disease | 65.9 (9.8, nr) | 57% | 24.4 (5.4, nr) | nr |

| <u>Reference</u> | <u>Measure</u> | <u>Language of measure</u> | <u>Measurement properties investigated</u> | <u>Residential status</u> | <u>N</u> | <u>Population</u> (N of each subgroup, or % where N not possible to calculate) | <u>Mean age</u> (SD, range) | <u>Gender</u> (% Male) | <u>Cognitive status</u> Mean MMSE (SD, range) unless otherwise stated | <u>Apathy score</u> Mean (SD, range) |
|------------------|----------------|----------------------------|--|---|------------------------|--|--|-----------------------------------|--|--|
| [99] | UPDRS | Norwegian [^] | Hypothesis Testing (convergent) | nr. Participants were assessed in outpatient clinics, at home and in nursing homes. No confirmation from correspondence regarding proportion of community-dwellers. | N=89 (convergent N=58) | Parkinson' Disease (41.4% with cognitive impairment) | 74.2 (8.8, nr) | 44.8% | 23.0 (7.2, nr) | UPDRS-apathy item nr. 17% had apathy according to diagnostic criteria. |
| [98] | UPDRS | English [^] | Hypothesis Testing (convergent) | Outpatients. Confirmed all community via correspondence with authors | N=301 | Parkinson's Disease | 67.8 (10.6; 30 to 90) | 63% | nr | 1.14 (1.1; 0 to 4) AS =13.7 (6.9) range =0 to 31. AS≥14: 50% |
| [105,106] | mds-UPDRS | English | Development (Item elicitation [including adaptation of items from UPDRS to create mds-UPDRS], Pilot study) | nr | nr | Item elicitation: nr. Pilot study: Part 1: Patients (PD, N=80), carers (N=nr) and professionals (N=nr) Part 2: Patients (N=32) and professionals (N=14) | nr | nr | nr | nr |
| [97] | mds-UPDRS | Hungarian | Hypothesis testing (convergent) | nr. Correspondence with author confirmed majority community. | N=584 | Parkinson's Disease PD with neurocognitive disorder (N=310) PD with depression (N=217) Apathy status: No apathy (N=477), Apathy (N=107) | <u>median (Q1 to Q3)</u> No apathy: 67 (61 to 73). Apathy: 68 (61 to 75) | No apathy: 60.2% Apathy: 52.3% | <u>median (Q1 to Q3)</u> No apathy: 28, (27 to 29) Apathy: 27 (24 to 28) | <u>median (Q1 to Q3)</u> LARS: No apathy: -26 (-30 to -21) Apathy: -15 (-22 to 5) |

45 Note: Where the study had used secondary data, the primary data sources were sought to gain the necessary information where it was not available in the article in question.

46 [^] Assumed based on location of study and/ or nationality of participants.

47 *Calculated by authors

48 Abbreviations: AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; ACE, Addenbrooke's Cognitive Examination; AD, Alzheimer's Disease; AES-12PD, Apathy Evaluation Scale 12-
49 item Parkinson's Disease; AES-C, Apathy Evaluation Scale Clinician; AES-I, Apathy Evaluation Scale Informant; AES-S, Apathy Evaluation Scale Self; AI, Apathy Inventory; AI-C, Apathy Inventory
50 Clinician; AI-I, Apathy Inventory Informant; ALS, Amyotrophic Lateral Sclerosis; AMI, Apathy Motivation Index; AS-S, Apathy Scale Self; AS-I, Apathy Scale Informant; bDAS, brief Dementia
51 Apathy Scale; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; CD, Current Depression; CDNOS, Cognitive Disorder Not Otherwise
52 Specified; Ctrl, Healthy Controls; DAIR, Dementia Apathy Interview Rating; DAS, Dementia Apathy Scale; DAS-I, Dementia Apathy Scale Informant; DAS-S, Dementia Apathy Scale Self; Dem,
53 Dementia; Dep, Depression; DEX, Dysexecutive Questionnaire; DLB, Dementia with Lewy Bodies; FrSBe-I, Frontal Systems Behavior Scale Informant; FtD, Frontotemporal Dementia; GDS,
54 Geriatric Depression Scale; GIP, Behavioral Rating Scale for Psychogeriatric Inpatients ; IMD, Index of Mental Decline; KBCI, Key Behaviors Change Inventory; LARS, Lille Apathy Rating Scale;
55 LARS-C, Lille Apathy Rating Scale Clinician; LARS-I, Lille Apathy Rating Scale Informant; LHS, Left Hemisphere Stroke; MCI, Mild Cognitive Impairment; MCS, Mild Cognitive Symptoms; MD,
56 Mixed Dementia; mds-UPDRS, Movement disorder Society Unified Parkinson's Disease Rating Scale; NPI, Neuropsychiatric Inventory; NPI-A, Neuropsychiatric Inventory Alternative; NPI-C,
57 Neuropsychiatric Inventory Clinician; nr, not reported; PD, Parkinson's Disease; PDa&d, Parkinson's Disease with apathy and depression; PDa, Parkinson's Disease with apathy; PDD,
58 Parkinson's Disease Dementia; PDd, Parkinson's Disease with depression; PDDd, Parkinson's Disease with dementia and depression; PDexclD, Parkinson's Disease without dementia;
59 PDexclDd, Parkinson's Disease without dementia or depression; PS, Parkinsonian Symptoms; RD, Remitted Depression; RHS, Right Hemisphere Stroke; SCD, Subjective Cognitive Decline;
60 UPDRS, Unified Parkinson's Disease Rating Scale; VaD, Vascular Dementia.

61

62

63 Table S.3. Risk of bias and results of development and content validity studies

| Reference | Measure | Met criteria? (Y/N) | Description | Relevance | | Comprehensiveness | | Comprehensibility | |
|-----------|---------|---------------------|--|-----------------------------|--|------------------------|--|------------------------|---|
| | | | | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [40,41] | AD-RD | Y | Development study: qualitative interviews for concept elicitation and expert review to refine the measure. | Apathy subscale: Inadequate | Construct of apathy is not clear. Items were all based on their mention by at least two carers (informal or formal) in qualitative interviews about how people with dementia express their mood. No justification was provided for the response options or recall period. (1?) | Doubtful | Patients or carers were not asked specifically about the measure. Expert review lead to reduction of items to avoid repetition. However, it was unclear what professionals were asked. (1?). | Doubtful | Patients or carers were not asked specifically about the measure. Expert review lead to modified instructions. However, it was unclear what professionals were asked (1?). |
| [33] | AES | Y | Development study and pilot study. | Inadequate | Construct of apathy is clear. Items were developed from the literature, professionals, and authors' observations and opinions of people with apathy, but participants not involved in eliciting items and observations not reported on. (1?). | | | Doubtful | Unclear what participants were asked. 14 items were removed from the preliminary item pool due to poor comprehensibility. (1?). |
| [34] | AI | Y | Development study. | Inadequate | Construct of apathy is clear. Items were developed from the literature and diagnostic criteria, but participants not involved in eliciting items. (1?). | | | | |
| [35] | AMI | N | Development study. | Inadequate | Construct of apathy is clear. Items were developed from the relevant items of the LARS and by professionals. Participants were not involved in eliciting items. (1?). | | | | |
| [36] | AS | Y | Development study (Adaptation of AES to make AS.) | Inadequate | Construct of apathy is clear. Participants not involved in eliciting items. Most relevant items of AES were selected by 2 professionals (S. Starkstein personal, communication, October 01, 2018). (1?). | Doubtful | Pilot study conducted with participants with neurological disorders, but not published, so unable to rate. New items were included by 2 professionals (S. Starkstein personal, communication, October 01, 2018). (1?). | Inadequate | Pilot study conducted with participants with neurological disorders, but not published, so unable to rate. Some items were modified by 2 professionals (S. Starkstein personal, communication, October 01, 2018). (1?). |
| [90] | AS-I | Y | Content validity study. | | | | | Doubtful | Unclear what participants were asked. Participants showed good understanding and no modifications were required (1?). |

| Reference | Measure | Met criteria? (Y/N) | Description | Relevance | | Comprehensiveness | | Comprehensibility | |
|-----------|-----------|---------------------|--|------------------------|---|------------------------|---|------------------------|--|
| | | | | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [42] | BMDS | Y | Development study. | Inadequate | Constructs of behaviour and mood, and apathy were not clear. Items were developed from the literature and author opinion, but participants not involved in eliciting items. (1?). | | | | |
| [43] | BSSD | Y | Development study and pilot study. | Inadequate | Items were developed from professionals and previous measures, but participants not involved in eliciting items. (1?). | Doubtful | Multiple pilot studies conducted to refine scale, but methods and results not reported. (1?). | Doubtful | Multiple pilot studies conducted to refine scale, but methods and results not reported. (1?). |
| [37] | DAIR | Y | Development study and pilot study. | Doubtful | Construct of apathy is clear. Items refer to apathy, and were developed with participation from people with dementia and carers. No justification was provided for the response options or recall period. (1+/-). | Doubtful | Unclear what participants were asked. (1?). | Doubtful | Unclear what participants were asked. (1?). |
| [38] | DAS | Y | Development study. | Inadequate | Items were developed from existing scales and experts, but participants not involved in eliciting items. (1?). | | | | |
| [68] | FrSBe-11a | Y | Content validity: cognitive interviewing study | | | | | Doubtful | 27% items had no discrepancies, with 82% of items having acceptable discrepancy*. However, participants do not appear to have been asked about the comprehensibility of instructions or response options. (1?) |
| [68] | FrSBe-14a | Y | Content validity: cognitive interviewing study | | | | | Doubtful | 21% items had no discrepancies, with 86% of items having acceptable discrepancy*. However, participants do not appear to have been asked about the comprehensibility of instructions or response options. (1?) |
| [46,47] | GDS | N | Development and pilot study (as a | Inadequate | Items were developed from professionals, but participants not involved in eliciting items. (1?). | | | Doubtful | Reported that patients accepted the measure, but methods by |

| Reference | Measure | Met criteria? (Y/N) | Description | Relevance | | Comprehensiveness | | Comprehensibility | |
|-----------|---------|------------------------|---------------------------------------|------------------------|--|------------------------|---|------------------------|---|
| | | | | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [49] | IMD | Y | measure of depression) Development | Inadequate | Items were developed from existing measures and professionals, but participants not involved in eliciting items. (1?). | | | | which this was ascertained were unclear. (1?) |
| [50,104] | KBCI | N | Development and pilot | Doubtful | Construct of apathy clear. Items were developed from the literature and interviews with patients, carers and professionals. Methods not clear. No justification for response options and recall period not clear. Patients and carers were later asked to rate the importance of items, and the majority were rated very or extremely important, but exact ratings not reported. (1+/-). | Doubtful | Patients and carers did not suggest any additional items. However, items were later removed after another phase in the development, so comprehensiveness may have changed. Method not clear. (1?) | Doubtful. | Patients and carers were asked about comprehensibility and no changes were suggested. Professionals were asked about comprehensibility and 15 items were re-worded. Methods and focus not clear (e.g. whether they were asked about each item, response options and recall period) (2?) |
| [39] | LARS | N | Development | Inadequate | Items were developed from Marin's concept of apathy and authors' clinical experience, but no systematic process and participants not involved in eliciting items. (1?). | | | | |
| [65] | LARS | Y | Pilot study | Doubtful | Participants asked about relevance, but results not reported. Methods and focus not clear (e.g. whether they were asked about each item, response options and recall period) (1?) | | | Doubtful | Participants asked about comprehensibility and format. Methods and focus not clear (e.g. whether they were asked about comprehensibility of instructions and response options as well as items) (1?) |
| [51] | NPI | N | Development and Delphi study | Inadequate | Items developed from the literature, but participants not involved in eliciting items. (1?). | Doubtful | Delphi panel of 10 professionals. Assessed "whether the essential elements of the behavior were captured" in each domain by rating screening and sub questions from 1 (well assessed) to 4 (poorly assessed). Apathy: screening | | |

| Reference | Measure | Met criteria? (Y/N) | Description | Relevance | | Comprehensiveness | | Comprehensibility | |
|-----------|-----------|---------------------|---|-----------------------------|---|------------------------|--|------------------------|---|
| | | | | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| | | | | | | | questions mean score = 1.3; sub-questions mean score = 1.4. No assessment of comprehensiveness by participants. (1?) | | |
| [87] | NPI-C | Y | Content validity (adaptation) | Doubtful | New items added from symptoms listed by alternative measures. Items were selected that were consistent with diagnostic criteria 2009. Participants not involved in eliciting new items. (1?) | Doubtful | Delphi panel of 8 professionals. Unclear what was asked. (1?) | Doubtful | Delphi panel of 8 professionals. Unclear what was asked. (1?) |
| [52] | UPDRS | N | Development study | Inadequate | Expert group elicited items from existing measures, but participants not involved in eliciting items. (1?). | | | Inadequate | Authors reviewed comprehensiveness of preliminary items. Changes were made and final version does not appear to have been reviewed. (1?) |
| [105,106] | mds-UPDRS | N | Development (Adaptation of UPDRS but involved new item elicitation and pilot study) | Apathy subscale: Inadequate | Expert group elicited items from literature, existing measures, clinical experience and participant survey, though methods not described in sufficient detail. Justification provided for response options but not recall period. (1?). | | | Doubtful | Comprehensiveness of preliminary items was reviewed by participants and professionals in a qualitative, then quantitative study. Items, instructions and response options were assessed. Unsure if recall period discussed. Changes were made in the first round and then again in the second round. (1?) |

64 Note: Studies only listed if they assessed content validity in some way or were a study describing the development of a measure. Some studies have multiple citations as multiple articles or
65 similar (e.g. PhD thesis) were published on the same study. Blank cells indicate this measurement property was not investigated by the study.

66 Quality of measurement property: Number of studies in parenthesis followed by rating: +, Sufficient; +/-, Inconsistent; -, Insufficient; ?, Indeterminate.

67 * Acceptable discrepancy was defined by the authors of the study as less than 30% of participants interpreting the items meaning in the way it was intended [68].

68 Abbreviations: AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; AES Apathy Evaluation Scale; AMI, Apathy Motivation Index; AI, Apathy Inventory; AS, Apathy Scale; AS-I,
69 Apathy Scale Informant; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; DAIR, Dementia Apathy Interview Rating; DAS, Dimensional
70 Apathy Scale; FrSBe-11a, Frontal Systems Behavior Scale 11 item apathy subscale; FrSBe-14a, Frontal Systems Behavior Scale 14 item apathy subscale; GDS, Geriatric Depression Scale; IMD,
71 Index of Mental Decline; KBCI, Key Behaviors Change Inventory; LARS, Lille Apathy Rating Scale; mds-UPDRS, Movement Disorder Society Unified Parkinson's Disease Rating Scale; NPI,
72 Neuropsychiatric Inventory; NPI-C, Neuropsychiatric Inventory Clinician; UPDRS, Unified Parkinson's Disease Rating Scale.

73 Unable to obtain development articles for: Dysexecutive Questionnaire (DEX), FrSBe and Behavioral Rating Scale for Psychogeriatric Inpatients (GIP).

74

75

77 Table S.4. Reviewer rating of content validity

| Measure | Relevance | | Comprehensiveness (quality rating) | Comprehensibility (quality rating) | Overall validity | |
|---------|---|---|---------------------------------------|--|--|--|
| | Older adults (quality rating) | Dementia & MCI (quality rating) | | | Older adults | Dementia & MCI |
| AD-RD | Unable to obtain the full list of items and instructions. | | | | | |
| AES | 94% relevant to apathy. 100% relevant to older adults. 94% relevant to research context. Response options appropriate. Suggested recall period too long, but personalised recall period also possible. (1+). | AES-I & AES-S: 94% relevant to apathy. 100% relevant to people with dementia. 94% relevant to research context. Response options appropriate. Suggested recall period too long, but personalised recall period also possible. (1+). AES-C: 94% relevant to apathy. 78% relevant to people with dementia, as some items based on where some items are rated based on patient free-recall. 94% relevant to research context. Response options appropriate. Suggested recall period too long, but personalised recall period also possible. (1+/-). | 3 domains of apathy included. (1+). | AES-I & AES-S: 94% appropriately worded. 72% match response options. (1+/-). AES-C: has additional guidance around this so AES-C response options deemed appropriate. (1+). | Sufficient (AES-I & AES-S: 2+, 1+/- ; AES-C: 3+) | Sufficient (2+, 1+/-) |
| AI | 100% items relevant to apathy, older adults and the research context. Response options appropriate for AI-C and AI-I, but not for AI-S. Recall period referencing onset of disease not appropriate for older adults, but personalised recall period possible. (Using the given recall period: 1+/-. Using the personalised recall period: 1+.) | 100% items relevant to apathy, people with dementia and the research context. Response options appropriate for AI-C and AI-I, but not for AI-S. Recall period of since onset of disease too long for people with dementia, but personalised recall period possible. (Using the given recall period: 1+/-. Using the personalised recall period: 1+.) | 3 domains of apathy included. (1+). | 0% of items appropriately worded. (1-) | Inconsistent (Given recall period: 1+, 1-, 1+/-; Personalised recall period: 2+, 1-) | Inconsistent (Given recall period: 1+, 1-, 1+/-; Personalised recall period: 2+, 1-) |
| AMI | 78% relevant to apathy. 100% relevant to older adults. 100% relevant to research context. Response options and recall period appropriate. (1+/-). | 78% relevant to apathy and to older adults. 100% relevant to research context. Response options and recall period appropriate. (1+/-). | 3 domains of apathy included. (1+). | 100% of items appropriately worded. 100% match response options. (1+). | Sufficient (2+, 1+/-) | Sufficient (2+, 1+/-) |

| Measure | Relevance | | Comprehensiveness (quality rating) | Comprehensibility (quality rating) | Overall validity | |
|---------|---|---|---------------------------------------|--|------------------------------|--------------------------------|
| | Older adults (quality rating) | Dementia & MCI (quality rating) | | | Older adults | Dementia & MCI |
| AS | 93% relevant to apathy. 93% relevant to older adults. 100% relevant to research context. Response options appropriate. Recall period too long. (1+). | 93% relevant to apathy. 100% relevant to people with dementia and the research context. Response options appropriate. Recall period too long. (1+) | 3 domains of apathy included. (1+). | 93% of items appropriately worded. 57% match response options (1+/-) | Sufficient (2+, 1+/-) | Sufficient (2+, 1+/-) |
| BMDS | 55% relevant to apathy. 100% relevant to older adults and the research context. Response options appropriate. Recall period uncertain. (1+/-). | 55% relevant to apathy. 100% relevant to people with dementia and the research context. Response options appropriate. Recall period uncertain. (1+/-). | Emotional dimension missing. (1-). | 100% of items appropriately worded, but combination with response options produces double negatives. (1+/-). | Inconsistent (1-, 2+/-) | Inconsistent (1-, 2+/-) |
| BSSD | 71% relevant to apathy. 100% relevant to older adults and research context. 14% response options appropriate. Recall period appropriate. (1+/-) | 71% relevant to apathy. 100% relevant to people with dementia and research context. 14% response options appropriate. Recall period appropriate. (1+/-) | 3 domains of apathy included. (1+). | 86% of items (questions directed at informants) appropriately worded. 100% match response options. (1+). | Sufficient (2+, 1+/-) | Sufficient (2+, 1+/-) |
| DAIR | 94% items relevant to apathy. 0% relevant for healthy older adults due to mandatory follow-up question relating to "illness". Response options appropriate. Recall period too long. (1+/-). | 94% items relevant to apathy. 100% relevant for people with dementia. Response options appropriate. Recall period too long. (1+). | 3 domains of apathy included. (1+). | 100% items appropriately worded. 81% match the response options. (1+/-). | Inconsistent (1+, 2+/-). | Sufficient (2+, 1+/-). |
| DAS | DAS: 79% items relevant to apathy. bDAS: 67% items relevant to apathy Both versions: 100% relevant to older adults. Response options appropriate. Recall period too long. (1+/-). | DAS: 79% items relevant to apathy. bDAS: 67% items relevant to apathy Both versions: 100% relevant to people with dementia. Response options appropriate. Recall period too long. (1+/-). | 3 domains of apathy included. (1+). | 100% of items appropriately worded. 100% match response options. (1+). | Sufficient (2+, 1+/-). | Sufficient (2+, 1+/-). |
| DEX | 63% items relevant to apathy. 100% relevant to older adults and research context. Complete response options not available. Recall period appropriate. (1+/-).* | 63% items relevant to apathy. 100% relevant to people with dementia and research context. Complete response options not available. Recall period appropriate. (1+/-).* | 3 domains of apathy included. (1+).* | Full wording not available, but 75% of items appear appropriately worded. Complete response options not known. (1?). | Inconsistent (1+, 1+/-, 1?)* | Inconsistent (1+/-, 1+/-, 1?)* |

| Measure | Relevance | | Comprehensiveness (quality rating) | Comprehensibility (quality rating) | Overall validity | |
|----------|---|--|--|---|------------------------------|------------------------------|
| | Older adults (quality rating) | Dementia & MCI (quality rating) | | | Older adults | Dementia & MCI |
| FrSBe | FrSBe-6a: 83% relevant to apathy. 100% relevant to older adults. FrSBe-11a: 82% relevant to apathy. 91% relevant to older adults. FrSBe-14a: 86% relevant to apathy. 93% relevant to older adults. And all versions: 100% relevant to research context. Response options not available. Recall period not appropriate for older adults. (1+/-).* | FrSBe-6a: 83% relevant to apathy. 100% relevant to older adults FrSBe-11a: 82% relevant to apathy. 91% relevant to people with mild dementia. FrSBe-14a: 86% relevant to apathy. 93% relevant to mild dementia. And all versions: 100% relevant to research context. Response options not available. Recall period not appropriate for people with dementia. (1+/-).* | All versions: 3 domains of apathy included. (1+).* | 6a: Full wording not available, but items suggests that 67% appropriately worded. Response options not available. (1?). 11a: Full wording not available, but items suggests that 91% appropriately worded. Response options not available. (1?). 14a: Full wording not available, but items suggests that 86% appropriately worded. Response options not available. (1?). | Inconsistent (1+, 1+/-, 1?)* | Inconsistent (1+, 1+/-, 1?)* |
| GDS-3a | 67% of items are relevant to apathy. All items relevant to older adults and the research context. Dichotomous response options not appropriate. Recall period appropriate. (1+/-). | 67% of items are relevant to apathy. All items relevant to people with dementia and the research context. Dichotomous response options not appropriate. Recall period appropriate. (1+/-). | Emotional dimension of apathy is missing (1-). | 100% appropriately worded and match response options. (1+). | Inconsistent (1+, 1-, 1+/-) | Inconsistent (1+, 1-, 1+/-) |
| GDS-6a | 50% of items are relevant to apathy. All items relevant to older adults and the research context. Dichotomous response options not appropriate. Recall period appropriate. (1+/-). | 50% of items are relevant to apathy. All items relevant to older adults and the research context. Dichotomous response options not appropriate. Recall period appropriate. (1+/-). | 3 domains of apathy included. (1+). | 100% appropriately worded and match response options. (1+). | Sufficient (2+, 1+/-) | Sufficient (2+, 1+/-) |
| GIP-9a | 44% of items relevant to apathy. 89% relevant to older adults in the community. 100% relevant to research context. Recall period appropriate. Response options not available. (1+/-).* | 44% of items relevant to apathy. 89% of items relevant to people with dementia in the community. 100% relevant to research context. . Recall period appropriate. Response options not available. (1+/-).* | Emotional dimension of apathy is missing. (1-).* | Full wording and official English translation of items not available, but authors translation suggest 89% appropriately worded. Response options not available. (1?).* | Inconsistent (1-, 1+/-, 1?) | Inconsistent (1-, 1+/-, 1?) |
| IMD | 100% of items relevant to apathy, older adults and the research context. Response options and recall period not available. (1?). | 100% of items relevant to apathy, people with dementia and the research context. Response options and recall period not available. (1?). | 3 domains of apathy included. (1+). | Full wording not available, but items suggest 33% appropriately worded. Response options not available. (1?). | Indeterminate (1+, 2?) | Indeterminate (1+, 2?) |
| KBCI-10a | 90% of items relevant to apathy. 80% of items relevant to older adults. All items relevant to research context. Response | 90% of items relevant to apathy. 80% of items relevant to people with dementia. All items relevant to research context. | 3 domains of apathy included. (1+). | 80% of items appropriately worded. 100% match response options. (1+/-). | Inconsistent (1+, 2+/-) | Inconsistent (1+, 2+/-) |

| Measure | Relevance | | Comprehensiveness (quality rating) | Comprehensibility (quality rating) | Overall validity | |
|----------------|--|--|--|---|---------------------------|---------------------------|
| | Older adults (quality rating) | Dementia & MCI (quality rating) | | | Older adults | Dementia & MCI |
| LARS | options appropriate. Recall period not available. (1+/-). 94% of items relevant to apathy. 100% relevant to older adults. Response options appropriate. Recall period too long. (1+). | Response options appropriate. Recall period not available. (1+/-). 94% of items relevant to apathy. 94% relevant to people with dementia. Response options appropriate. Recall period too long. (1+). | 3 domains of apathy included. (1+). | 87% appropriately worded. 100% match response options. (1+). | Sufficient (3+) | Sufficient (3+) |
| NPI (original) | 100% of items relevant to apathy, older adults and the research context. Response options appropriate. Suggested recall period too long, but personalised recall period also possible. (1+). | 100% of items relevant to apathy, people with dementia and the research context. Response options appropriate. Suggested recall period too long, but personalised recall period also possible. (1+). | Emotional dimension of apathy is missing from the screening questions. No dimensions are rated separately. (1-). | Assessments of frequency and severity are based on multiple symptoms, so could be considered a double barrelled question and therefore not appropriately worded. However carers are advised to rate the worst one. 100% match the response options. (1+). | Inconsistent (2+, 1-) | Inconsistent (2+, 1-) |
| NPI-A | Unable to obtain full instructions and guidance. | | | | | |
| NPI-C | 100% of items relevant to apathy, older adults and the research context. Response options appropriate. Recall period too long. (1+). | 100% of items relevant to apathy, people with dementia and the research context. Response options appropriate. Recall period too long. (1+). | 3 domains of apathy included. (1+). | Assessments of frequency and severity are based on multiple symptoms, so could be considered a double barrelled question and therefore not appropriately worded. However carers are advised to rate the worst one. 100% match the response options. (1+). | Sufficient (3+) | Sufficient (3+) |
| UPDRS | 100% relevant to apathy, older adults and research context. (Note: only 1 item). Response options appropriate. Recall period not clear. (1+). | 100% relevant to apathy, people with dementia and research context. (Note: only 1 item). Response options appropriate. Recall period not clear. (1+). | Emotional domain of apathy missing. Cognitive and Behavioural elements included but not rated separately. (1-). | Item wording is not given, or could not be obtained; only the heading is provided, so it is unclear if it matches the response options. (1?). | Inconsistent (1+, 1-, 1?) | Inconsistent (1+, 1-, 1?) |
| mds-UPDRS | 100% relevant to apathy, older adults and the research context. (Note: only 1 item). Response options and recall period appropriate. (1+). | 100% relevant to apathy, people with dementia and the research context. (Note: only 1 item). Response options and recall period appropriate. (1+). | Emotional domain of apathy missing. Cognitive and Behavioural elements included but not rated separately. (1-). | 100% appropriate worded and match response options. (1+). | Inconsistent (2+, 1-) | Inconsistent (2+, 1-) |

78 *based on list of apathy items presented by another publication (DEX [81]; FrsBE [64,68]; GIP [107])

79 Quality of measurement property: Number of studies in parenthesis followed by rating: +, Sufficient; +/-, Inconsistent; -, Insufficient; ?, Indeterminate.

80 Abbreviations: AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; AES Apathy Evaluation Scale; AES-C, Apathy Evaluation Scale Clinician; AES-I, Apathy Evaluation Scale
81 Informant; AES-S, Apathy Evaluation Scale Self; AI, Apathy Inventory; AI-C, Apathy Inventory Clinician; AI-I, Apathy Inventory Informant; AI-S, Apathy Inventory Self; AMI, Apathy Motivation
82 Index; AS, Apathy Scale; b-DAS, brief-Dimensional Apathy Scale; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; DAIR, Dementia Apathy
83 Interview Rating; DAS, Dimensional Apathy Scale; DEX, Dysexecutive Questionnaire; FrSBe, Frontal Systems Behavior Scale; FrSBe-6a, Frontal Systems Behavior Scale 6-item apathy subscale;
84 FrSBe-11a, Frontal Systems Behavior Scale 11-item apathy subscale; FrSBe-14a, Frontal Systems Behavior Scale 14-item apathy subscale; GDS-3a, Geriatric Depression Scale 3 item apathy
85 subscale; GDS-6a, Geriatric Depression Scale 6 item apathy subscale; GIP-9a, Behavioral Rating Scale for Psychogeriatric Inpatients 9 item apathy subscale; IMD, Index of Mental Decline;
86 KBCI-10a, Key Behaviors Change Inventory 10 item apathy subscale; LARS, Lille Apathy Rating Scale; MDS-UPDRS, Movement Disorder Society-Sponsored Revision of the Unified Parkinson's
87 Disease Rating Scale; NPI, Neuropsychiatric Inventory; NPI-A, Neuropsychiatric Inventory Alternative; NPI-C, Neuropsychiatric Inventory Clinician; UPDRS, Unified Parkinson's Disease Rating
88 Scale

89

90

92 Table S.5. Risk of bias and results of studies of remaining measurement properties

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|----------|------------------------|---|------------------------|---------------------------|-------------------------|------------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [40] | AD-RD | | | | | 1 Doubtful. | r=.72 (1+). | | | | |
| [78] | AES-12PD | | | 3 Very good. | α =.90 to .92 (3+) | | | | | 3 Adequate. 1 Very Good. | 3 met hypothesis (3+). 1 did not meet hypothesis (1-). |
| [53] | AES-C | 1 Doubtful. | 3 factors (57.06%): Apathy (40.02%); Novelty Seeking (9.35%); Insight & social (7.68%). [1+] | 1 Very good. | α =.90. (1+). | 2 Doubtful. | r=.88 to .86 (2+). | | | 2 Inadequate. 2 Very Good. | 4 met hypothesis (4+). |
| [54] | AES-C | 1 Doubtful. | 2 factors (51.1%): Apathy (42.4%); Interest (8.7%). [1+] | | | | | | | 1 Inadequate. 1 Doubtful. 2 Adequate. | 2 met hypothesis (2+). 2 did not meet hypothesis (2-). |
| [55] | AES-C | 1 Inadequate. | 3 factors (84.17 [^]): Interest & Motivation (39.72% [^]); Task Completion (29.67% [^]); Insight (14.78% [^]). [1-] | 1 Very good. | α =.93. (1+). | | | | | 1 Inadequate. 1 Doubtful. 1 Very Good. | 3 met hypothesis (3+). |
| [33] | AES-C | 1 Inadequate. | 3 factors: Apathy (32-53%); Novelty Seeking (5-10%); Insight & dependency (7-8%). [1?] | 1 Very good. | α =.90. (1+). | 1 Doubtful. 1 Adequate. | r=.88 (1+). ICC= .94 (+). | | | 3 Inadequate. 1 Doubtful. 1 Adequate. 4 Very Good. | 5 met hypothesis (5+). 1 did not meet hypothesis (1-). 3 insufficient information (3?). |
| [54] | AES-I | 1 Doubtful. | 2 factors (54.4%): Interest (45.1%); Apathy (9.3%). [1+] | | | | | | | 1 Inadequate. 1 Doubtful. 2 Adequate. | 2 met hypothesis (2+). 2 did not meet hypothesis (2-). |
| [55] | AES-I | | | 1 Very good. | α =.89. (1+). | | | | | 1 Inadequate. 1 Doubtful. 1 Very Good. | 3 met hypothesis (3+). |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|----------|-------------------------|---|------------------------|-------------------------|------------------------|-------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [33] | AES-I | 1 Inadequate. | 3 factors: Apathy (32-53%); Novelty Seeking (5-10%); Insight & dependency (7-8%). [1?] | 1 Very good. | $\alpha=.94.$ (1+). | 1 Doubtful. | $r=.94$ (1+). | | | 3 Inadequate. 1 Doubtful. 1 Adequate. 4 Very Good. | 4 met hypothesis (4+). 2 did not meet hypothesis (2-). 3 insufficient information (3?). |
| [56] | AES-I | 1 Doubtful. 1 Adequate. | 2 factors (62.56%^): Factor 1 (56.2%); Factor 2 (6.36%). [1+]. 1 factor (62.8%). [1+]. | 1 Very good. | $\alpha=.95.$ (1+). | | | n/a | SEM=2.9. (1?). | 2 Very Good. | 2 met hypothesis (2+). |
| [89] | AES-I | | | | | | | | | 1 Adequate. 2 Very good. | 1 met hypothesis (1+). 2 did not meet hypothesis (2-). |
| [71] | AES-I | | | 1 Very good. | $\alpha=.88.$ (1+). | | | | | | |
| [71] | AES-I-16 | | | 1 Very good. | $\alpha=.90.$ (1+). | | | | | 1 Adequate. 1 Very Good | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). |
| [55] | AES-S | | | 1 Very good. | $\alpha=.90.$ (1+). | | | | | 1 Inadequate. 1 Doubtful. 1 Very Good. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). 1 insufficient information (1?) |
| [54] | AES-S | 1 Doubtful. | 2 factors (43.3%^): Apathy (36.4%); Other (6.9%) [1+] | | | | | | | 1 Inadequate. 1 Doubtful. 2 Adequate. | 2 met hypothesis (2+). 2 did not meet hypothesis (2-). |
| [33] | AES-S | 1 Inadequate. | 3 factors: Apathy (32-53%); Novelty Seeking (5-10%); Insight & dependency (7-8%). [1?] | 1 Very good. | $\alpha=.86.$ (1+). | 1 Doubtful. | $r=.76$ (1+). | | | 3 Inadequate. 1 Doubtful. 1 Adequate. 4 Very Good. | 5 met hypothesis (5+). 1 did not meet hypothesis (1-). 3 insufficient information (3?) |
| [56] | AES-S | 1 Doubtful. 1 Adequate. | 2 factors (61.69%^): Factor 1 (55.37%); Factor 2 (6.32%). [1+]. 1 factor (61.2%). [1+]. | 1 Very good. | $\alpha=.95.$ (1+). | | | n/a | SEM=2.7. (1?). | 2 Very Good. | 2 met hypothesis (2+). |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|---------|------------------------|--|------------------------|---------------------------------|---------------------------|-------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [89] | AES-S | | | | | | | | | 1 Adequate. 2 Very good. | 1 met hypothesis (1+). 2 did not meet hypothesis (2-). |
| [57] | AES-S | 2 Doubtful. | 3 factors (58%): Apathy (38.27%); Friendship (10.86%); Other (8.88%) [1+]. 3 factors (59.54%; variance explained per factor not reported.) [1?] | 2 Very good. | $\alpha=.90$ to .92. (2+). | | | | | 2 Doubtful. 4 Adequate. 4 Very Good. | 5 met hypothesis (5+). 3 did not meet hypothesis (3-). |
| [80] | AI-C | | | | | 1 Doubtful. | ICC=.97 (1+). | | | 1 Inadequate. | 1 met hypothesis (1+). |
| [72] | AI-C | | | 1 Doubtful. | $\alpha=.83$. (1?). | | | | | | |
| [34] | AI-I | | | 1 Doubtful. | $\alpha=.84$. (1?). | 1 Doubtful. 1 Inadequate. | Kappa= .96 to .99 (2+). | | | 1 Adequate. 3 Very Good. | 3 met hypothesis (3+). 1 did not meet hypothesis (1-). |
| [72] | AI-I | | | 1 Doubtful. | $\alpha=.83$. (1?). | | | | | | |
| [34] | AI-S | | | | | | | | | 3 Very Good. | 1 met hypothesis. (1+). 2 did not (2-) |
| [72] | AI-S | | | 1 Doubtful. | $\alpha=.61$. (1?). | | | | | | |
| [67] | AMI | | | * | $\alpha=.86$ | | | | | 2 Adequate. | 2 did not meet hypothesis (2-). |
| [58] | AS-HC | 1 Very Good. | 1 factor CFI=1.00, RMSEA=0.00. [1+] | 1 Very Good | $\alpha=.94$. (1+). | | | | | 1 Very Good. | 1 did not meet hypothesis (1-). |
| [90] | AS-I | | | | | | | | | 1 Inadequate. 1 Doubtful. 1 Very Good. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). 1 insufficient information available (1?). |
| [59] | AS-S | 1 Doubtful. | 13-item: 3 factors (55.61%). Variance explained per factor not reported. [1?] | 2 Doubtful. | 14 item version: $\alpha=.82$. | | | | | | |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|---------|--------------------------|--|------------------------|---|---------------------------|------------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [36] | AS-S | | | 1 Doubtful. | 13 item version: $\alpha=.85$. (2?). $\alpha=.76$. (1?). | 2 Doubtful. | $r=.81$ to $.90$. (2+). | | | 1 Doubtful. | 1 met hypothesis (1+). |
| [58] | AS-S | 1 Very Good. | 1 factor. CFI=1.00, RMSEA=0.00. [1+]. | | | | | | | | |
| [60] | AS-S | 2 Adequate. | 14-item: 2 factors (57.7%): Cognitive-Behavioural (24.2%); Apathy and insight (15.05%). [1-]. 13-item: 2 factors (41.7%) Variance explained per factor not reported. [1?] | 2 Doubtful. | 14 item: $\alpha=.69$. 13 item: $\alpha=.74$. (2?). | | | | | 1 Adequate. 2 Very Good. | 3 met hypothesis (3+). |
| [77] | AS-S | | | 1 Inadequate. | Guttman's $\lambda = .89$. (1?). | 1 Inadequate. | ICC=.78 (1+). n/a | SEM = 2.34. (1?). | | 1 Doubtful. 2 Very Good. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). 1 insufficient information (1?). |
| [61] | AS-S | 1 Very Good. 1 Adequate. | AS-S: 3 factors (nr). [1+/-]. 11 item: 2 factors: 54.1% of variance explained. [1-]. | 11-item: 1 Inadequate. | 11 item: $\alpha=.77$ (1?) | | | | | | |
| [91] | AS-S | | | 1 Doubtful. | $\alpha=.78$. (1?). | | | | | 1 Inadequate. 2 Doubtful. 1 Very good. | 3 met hypothesis (3+). 1 did not meet hypothesis (1-). |
| [42] | BMDS | | | | | 1 Inadequate. | $r=.90$. (1+) | | | | |
| [43] | BSSD | | | 1 Doubtful. | $\alpha=.82$ to $.83$ (1?) | 1 Inadequate. 3 Doubtful. | ICC=.65 to $.85$. (2+, 2-). | | | 2 Inadequate. 1 Doubtful. 1 Very Good. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|-----------|------------------------|--|------------------------|-------------------------|------------------------|-------------------------|------------------------|-------------------------|--|--|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [37] | DAIR | 1 Adequate. | 1 factor (38%) [1+] | 1 Very Good. | $\alpha=.89$. (1+). | 1 Inadequate. | $r=.85$ (1+) | 1 Doubtful. | 100% agreement (1+). | 2 Inadequate. 2 Very Good. | (1-). 2 insufficient information (2?). 3 met hypothesis (3+). 1 did not meet hypothesis (1+). |
| [75] | DAS-I | | | * | $\alpha=.92$ | | | | | 2 Adequate. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). |
| [74] | DAS-I | | | * | $\alpha=.93$ | | | | | 2 Adequate. | 2 met hypothesis (2+). |
| [75] | DAS-S | | | * | $\alpha=.84$ | | | | | 2 Adequate. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). |
| [74] | DAS-S | | | * | $\alpha=.85$ | | | | | 2 Adequate. | 2 met hypothesis (2+). |
| [62] | DAS-S | * | 3 factors (45.87%) Organisation & planning (28.21%); Initiation (9.76%); Emotional (7.90%). | * | $\alpha=.87$ | | | | | 4 Adequate. 2 Very Good. | 4 met hypothesis (4+). 2 did not meet hypothesis (2-). |
| [63] | bDAS | * | Item Hi=.40 to .76. No other fit measures reported. | | | | | | | | |
| [76] | bDAS | | | * | $\alpha=.81$. | 1 Inadequate. | ICC=.84 (1+). | | | | |
| [81] | DEX | | | | | 1 Doubtful. | ICC=.93 (1+). | | | 1 Inadequate. 1 Adequate. 2 Very Good. | 2 met hypothesis (2+). 2 did not meet hypothesis (2+) |
| [64] | FrSBe-6a | | | 1 Doubtful. | $\alpha=.88$. (1?). | | | | | | |
| [68] | FrSBe-11a | | | 1 Doubtful. | $\alpha=.83$. (1?). | | | | | | |
| [68] | FrSBe-14a | | | 1 Doubtful. | $\alpha=.88$. (1?). | | | | | 6 Doubtful. | 5 met hypothesis (5+). 1 did not meet hypothesis (1-). |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|----------------------------|------------------------|--|------------------------|-------------------------|---------------------------|-------------------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [64] | FrSBe-14a | 1 Inadequate. | 1 Factor specified: 12 out of 14 items had loadings >.40. (nr). [1?]. | 1 Doubtful. | $\alpha=.80$. (1?). | | | | | | |
| [92] | GDS-3a | | | | | | | | | 2 Adequate. | 2 did not meet hypothesis. (2+). |
| [69] | GDS-6a | | | 1 Doubtful. | $\alpha=.51$ (1?). | | | | | 1 Doubtful. 2 Adequate. | 3 met hypothesis (3+). |
| [82] | GIP- apathy subscale | | | | | 1 Doubtful. | ICC=.72 (1+). | n/a | SEM=1.22. (1?) | | |
| [82] | GIP- apathy domain | | | | | 1 Doubtful. | ICC=.83 (1+). | n/a | SEM=1.38. (1?) | | |
| [49] | IMD | | | | | | | | | 1 Inadequate. 3 Doubtful. | 3 met hypothesis (3+). 1 insufficient information (1?). |
| [93] | KBCI | | | | | | | | | 1 Inadequate. 1 Doubtful. 5 Adequate. | 6 met hypothesis (6+). 1 did not meet hypothesis (1-). |
| [83] | LARS-C | | | | | 2 Doubtful. | ICC=.94 to .99 (2+). | | | 2 Inadequate. 2 Doubtful. 5 Very Good. | 7 met hypothesis (7+). 1 did not meet hypothesis (1-). 1 insufficient information (1?). |
| [65] | LARS-C | * | 4 factors (67.5%): intellectual curiosity (nr); emotion (nr); action-initiation (nr); self awareness (nr). | * | $\alpha=.81$. (*) | 2 Doubtful. | ICC= .97. (1+). Kappa = .93 (1+). | | | 1 Inadequate. 2 Adequate. | 2 met hypothesis (2+). 1 did not meet hypothesis (1-). |
| [70] | LARS-I | | | * | $\alpha=.87$. (*) | 2 Doubtful. | ICC =.99. (1+). (1+)ICC =.99. (1+). | | | 2 Adequate. | 2 met hypothesis (2+). |
| [84] | NPI | | | | | 1 Doubtful. 1 Inadequate. | ICC = .67 (1-). $r_s= .53$ (1-). | | | | |

| Reference | Measure | Structural validity | | Internal consistency | | Reliability | | Measurement error | | Hypothesis testing | |
|-----------|-----------|------------------------|--|------------------------|-------------------------|---------------------------|-------------------------------|------------------------|-------------------------|--|---|
| | | Methodological quality | Result (% variance explained) [quality rating] | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) | Methodological quality | Result (quality rating) |
| [94] | NPI | | | | | | | | | 2 Doubtful. | 2 insufficient information (2?). |
| [51] | NPI | | | | | | | | | 1 Doubtful. | 1 insufficient information (1?). |
| [85] | NPI | | | | | 1 Inadequate | r=.96 (1+). | | | 1 Doubtful. | 1 insufficient information (1?). |
| [73] | NPI | | | 1 Doubtful. | α =.82 (1?) | 1 Doubtful. 1 Inadequate. | ICC=.87. (1+). r=.76 (1+). | | | 1 Inadequate. 1 Doubtful. 1 Very Good. | 1 met hypothesis (1+). 1 did not meet hypothesis (1-). 1 insufficient information (1?). |
| [95] | NPI | | | | | | | | | 1 Inadequate. | 1 did not meet hypothesis (1-). |
| [79] | NPI | | | 1 Doubtful. | α =.83 (1?) | 1 Doubtful. | Kendell CC=1.00 (1+). | | | 1 Inadequate. | 1 did not meet hypothesis (1-). |
| [96] | NPI | | | | | | | | | 1 Very Good. | 1 did not meet hypothesis (1-). |
| [86] | NPI | | | | | 1 Doubtful. | ICC=.99 (1+). | | | | |
| [66] | NPI-A | 1 Adequate. | 1 factor (66%). [1+]. | 1 Very Good. | α =.91 (1+) | | | | | | |
| [87] | NPI-C | | | | | 1 Doubtful. | Item ICC= .74 to .89 (1+). | | | 1 Adequate. | 1 did not meet hypothesis (1-). |
| [88] | NPI-C | | | | | 1 Doubtful. | ICC=.87 (1+). | | | 1 Adequate. | 1 met hypothesis (1+). |
| [97] | mds-UPDRS | | | | | | | | | 1 Very Good. | 1 met hypothesis (1+). |
| [98] | UPDRS | | | | | | | | | 1 Adequate. | 1 met hypothesis (1+). |
| [99] | UPDRS | | | | | | | | | 1 Inadequate. 2 Very Good. | 3 did not meet hypothesis (3-). |
| [100] | UPDRS | | | | | | | | | 1 Very Good. | 1 met hypothesis (1+). |

93
94
95
96

Blank cells indicate this measurement property was not investigated.

*Was assessed by the study, but methodological quality rating nor quality rating of result conducted, as the measure is based on a formative model.

^ Value calculated by review team based on information provided in the article.

Quality of measurement property: Number of studies in parenthesis followed by rating: +, Sufficient; +/-, Inconsistent; -, Insufficient; ?, Indeterminate.

97 Abbreviations: +, Sufficient; -, Insufficient; ?, Indeterminate; AD-RD, Alzheimer's Disease and Related Dementias Mood Scale; AES-12PD, Apathy Evaluation Scale for Parkinson Disease; AES-C,
98 Apathy Evaluation Scale Clinician; AES-I, Apathy Evaluation Scale Informant; AES-I-16, Apathy Evaluation Scale Informant 16 item version; AES-S, Apathy Evaluation Scale Self; AI-C, Apathy
99 Inventory Clinician; AI-I, Apathy Inventory Informant; AI-S, Apathy Inventory Self; AMI, Apathy Motivation Index; AS-HC, Apathy Scale Home Care; AS-I, Apathy Scale Informant; AS-S, Apathy
100 Scale Self; b-DAS, brief-Dimensional Apathy Scale; BMDS, Behavioural and Mood Disturbance Scale; BSSD, Behavioral Syndromes Scale for Dementia; DAIR, Dementia Apathy Interview
101 Rating; DAS-I, Dimensional Apathy Scale Informant; DAS-S, Dimensional Apathy Scale Self; DEX, Dysexecutive Questionnaire; FrSBe-6a, Frontal Systems Behavior Scale 6-item apathy
102 subscale; FrSBe-11a, Frontal Systems Behavior Scale 11-item apathy subscale; FrSBe-14a, Frontal Systems Behavior Scale 14-item apathy subscale; GDS-3a, Geriatric Depression Scale 3 item
103 apathy subscale; GDS-6a, Geriatric Depression Scale 6 item apathy subscale GIP, Behavioral Rating Scale for Psychogeriatric Inpatients; IMD, Index of Mental Decline; KBCI, Key Behaviors
104 Change Inventory; LARS-C, Lille Apathy Rating Scale Clinician; LARS-I, Lille Apathy Rating Scale Informant; mds-UPDRS, Movement Disorder Society-Sponsored Revision of the Unified
105 Parkinson's Disease Rating Scale; NPI, Neuropsychiatric Inventory; NPI-A, Neuropsychiatric Inventory Alternative; NPI-C, Neuropsychiatric Inventory Clinician; nr, not reported; UPDRS,
106 Unified Parkinson's Disease Rating Scale
107 Where there is no rating available for the researcher, this means it was not possible to obtain sufficient information regarding the measure to assess its content validity. Ratings of content
108 validity are for both people with dementia or MCI and older adults unless otherwise specified.

109

110 **References**

- 111 1 Robert PH, Lanctôt KL, Agüera-Ortiz L, Aalten P, Bremond F, Defrancesco M, et al. Is it time to revise the diagnostic criteria for apathy in brain disorders?
112 The 2018 international consensus group. *European Psychiatry*. 2018 Oct;54:71–6.
- 113 2 Bergh S, Selbæk G. The prevalence and the course of neuropsychiatric symptoms in patients with dementia. 1. 2012 Nov;22(2). DOI:
114 10.5324/nje.v22i2.1570
- 115 3 Cipriani G, Lucetti C, Danti S, Nuti A. Apathy and dementia. Nosology, assessment and management. *J Nerv Ment Dis*. 2014 Oct;202(10):718–24.
- 116 4 Monastero R, Mangialasche F, Camarda C, Ercolani S, Camarda R. A Systematic Review of Neuropsychiatric Symptoms in Mild Cognitive Impairment.
117 *Journal of Alzheimer's Disease*. 2009 Jan;18(1):11–30.
- 118 5 Martin E, Velayudhan L. Neuropsychiatric Symptoms in Mild Cognitive Impairment: A Literature Review. *DEM*. 2020;49(2):146–55.
- 119 6 Lam LCW, Tam CWC, Chiu HFK, Lui VWC. Depression and apathy affect functioning in community active subjects with questionable dementia and mild
120 Alzheimer's disease. *Int J Geriatr Psychiatry*. 2007 May;22(5):431–7.

- 121 7 Benoit M, Andrieu S, Lechowski L, Gillette-Guyonnet S, Robert PH, Vellas B, et al. Apathy and depression in Alzheimer's disease are associated with
122 functional deficit and psychotropic prescription. *Int J Geriatr Psychiatry*. 2008 Apr;23(4):409–14.
- 123 8 Mukherjee A, Biswas A, Roy A, Biswas S, Gangopadhyay G, Das SK. Behavioural and Psychological Symptoms of Dementia: Correlates and Impact on
124 Caregiver Distress. *DEE*. 2017;7(3):354–65.
- 125 9 Sultzer DL. Why Apathy in Alzheimer's Matters. *Am J Psychiatry*. 2018 01;175(2):99–100.
- 126 10 Armstrong N, Schupf N, Grafman J, Huey ED. Caregiver Burden in Frontotemporal Degeneration and Corticobasal Syndrome. *DEM*. 2013;36(5–6):310–8.
- 127 11 Ready RE, Ott BR, Grace J, Cahn-Weiner DA. Apathy and executive dysfunction in mild cognitive impairment and Alzheimer disease. *Am J Geriatr Psychiatry*.
128 2003 Apr;11(2):222–8.
- 129 12 Stuss DT, Van Reekum R, Murphy KJ. Differentiation of states and causes of apathy. *The neuropsychology of emotion*. New York, NY, US: Oxford University
130 Press; 2000; pp 340–63.
- 131 13 Yeager CA, Hyer L. Apathy in dementia: relations with depression, functional competence, and quality of life. *Psychol Rep*. 2008 Jun;102(3):718–22.
- 132 14 Pickett J, Bird C, Ballard C, Banerjee S, Brayne C, Cowan K, et al. A roadmap to advance dementia research in prevention, diagnosis, intervention, and care
133 by 2025. *International journal of geriatric psychiatry*. 2018 Jul;33(7):900–6.
- 134 15 Onyike CU, Sheppard J-ME, Tschanz JT, Norton MC, Green RC, Steinberg M, et al. Epidemiology of apathy in older adults: the Cache County Study. *Am J*
135 *Geriatr Psychiatry*. 2007 May;15(5):365–75.
- 136 16 Chilovi BV, Conti M, Zanetti M, Mazzù I, Rozzini L, Padovani A. Differential Impact of Apathy and Depression in the Development of Dementia in Mild
137 Cognitive Impairment Patients. *DEM*. 2009;27(4):390–8.
- 138 17 Starkstein SE, Jorge R, Mizrahi R, Robinson RG. A prospective longitudinal study of apathy in Alzheimer's disease. *J Neurol Neurosurg Psychiatr*. 2006
139 Jan;77(1):8–11.
- 140 18 Wadsworth LP, Lorus N, Donovan NJ, Locascio JJ, Rentz DM, Johnson KA, et al. Neuropsychiatric Symptoms and Global Functional Impairment along the
141 Alzheimer's Continuum. *DEM*. 2012;34(2):96–111.

- 142 19 Gilmore-Bykovskiy A. Commentary on Apathy as a Model for Investigating Behavioral and Psychological Symptoms in Dementia. *Journal of the American*
143 *Geriatrics Society*. 2018;66(S1):S13–6.
- 144 20 Clarke DE, Ko JY, Kuhl EA, van Reekum R, Salvador R, Marin RS. Are the available apathy measures reliable and valid? A review of the psychometric
145 evidence. *Journal of psychosomatic research*. 2011;70(1):73–97.
- 146 21 Radakovic R, Harley C, Abrahams S, Starr JM. A systematic review of the validity and reliability of apathy scales in neurodegenerative conditions. *Int*
147 *Psychogeriatr*. 2015;27(6):903–23.
- 148 22 Mohammad D, Ellis C, Rau A, Rosenberg PB, Mintzer J, Ruthirakuhan M, et al. Psychometric Properties of Apathy Scales in Dementia: A Systematic Review. *J*
149 *Alzheimers Dis*. 2018;66(3):1065–82.
- 150 23 Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy
151 included in systematic reviews. *BMC Medical Research Methodology*. 2003 Nov;3(1):25.
- 152 24 Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome
153 measures. *Quality of Life Research*. 2018 Feb DOI: 10.1007/s11136-018-1798-3
- 154 25 Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-
155 reported outcome measures: a Delphi study. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*.
156 2018 Mar DOI: 10.1007/s11136-018-1829-0
- 157 26 Burgon C, Goldberg S, van der Wardt V, Harwood RH. A systematic review of measures of apathy for older adults: validity, reliability and conceptualisation
158 of apathy. *PrAISED Discussion Paper Series*. 2019 Jul;(5).
- 159 27 Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported
160 Outcome Measures. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*. 2017 Dec DOI:
161 10.1007/s11136-017-1765-4
- 162 28 de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge: Cambridge University Press; 2011. DOI:
163 10.1017/CBO9780511996214

- 164 29 Park MS, Kang KJ, Jang SJ, Lee JY, Chang SJ. Evaluating test-retest reliability in patient-reported outcome measures for older people: A systematic review.
165 International Journal of Nursing Studies. 2018 Mar;79:58–69.
- 166 30 Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of
167 measurement instruments. Quality of Life Research. 2009 27 01/23/received 08/06/accepted;18(8):1115–23.
- 168 31 Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology,
169 and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737–45.
- 170 32 Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a “Core
171 Outcome Set” – a practical guideline. Trials. 2016 Sep;17(1):449.
- 172 33 Marin RS, Biedrzycki RC, Firinciogullari S. Reliability and validity of the Apathy Evaluation Scale. Psychiatry Res. 1991 Aug;38(2):143–62.
- 173 34 Robert PH, Clairet S, Benoit M, Koutaich J, Bertogliati C, Tible O, et al. The apathy inventory: assessment of apathy and awareness in Alzheimer’s disease,
174 Parkinson’s disease and mild cognitive impairment. Int J Geriatr Psychiatry. 2002 Dec;17(12):1099–105.
- 175 35 Ang Y-S, Lockwood P, Apps MAJ, Muhammed K, Husain M. Distinct Subtypes of Apathy Revealed by the Apathy Motivation Index. PLOS ONE. 2017;12(1):1–
176 15.
- 177 36 Starkstein SE, Mayberg HS, Preziosi TJ, Andrezejewski P, Leiguarda R, Robinson RG. Reliability, validity, and clinical correlates of apathy in Parkinson’s
178 disease. J Neuropsychiatry Clin Neurosci. 1992;4(2):134–9.
- 179 37 Strauss ME, Sperry SD. An informant-based assessment of apathy in Alzheimer disease. Neuropsychiatry Neuropsychol Behav Neurol. 2002 Sep;15(3):176–
180 83.
- 181 38 Radakovic R, Abrahams S. Developing a new apathy measurement scale: Dimensional Apathy Scale. Psychiatry Research. 2014 Nov;219(3):658–63.
- 182 39 Sockeel P, Dujardin K, Devos D, Denève C, Destée A, Defebvre L. The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy:
183 validation in Parkinson’s disease. J Neurol Neurosurg Psychiatry. 2006 May;77(5):579–84.
- 184 40 Tappen RM, Williams CL. Development and Testing of the Alzheimer’s Disease and Related Dementias Mood Scale. Nurs Res. 2008;57(6):426–35.

- 185 41 Tappen RM, Williams C. Attribution of emotion in advanced Alzheimer's disease: Family and caregiver perspectives. *American Journal of Alzheimer's*
186 *Disease*. 1998 Sep;13(5):257–64.
- 187 42 Greene JG, Smith R, Gardiner M, Timbury GC. Measuring behavioural disturbance of elderly demented patients in the community and its effects on
188 relatives: a factor analytic study. *Age Ageing*. 1982 May;11(2):121–6.
- 189 43 Devanand DP, Brockington CD, Moody BJ, Brown RP, Mayeux R, Endicott J, et al. Behavioral syndromes in Alzheimer's disease. *Int Psychogeriatr*. 1992;4
190 Suppl 2:161–84.
- 191 44 Burgess PW, Alderman N, Wilson BA, Evans JJ, Emslie H. The Dysexecutive Questionnaire. Behavioural assessment of the dysexecutive syndrome. Bury St
192 Edmunds, UK: Thames Valley Test Company; 1996.
- 193 45 Grace J, Malloy PF. *Frontal Systems Behavior Scale (FrSBe): Professional manual*. Lutz, FL: Psychological Assessment Resources; 2001.
- 194 46 Brink TL, Yesavage JA, Lum O, Heersema PH, Adey M, Rose TL. Screening tests for geriatric depression. *Clinical Gerontologist: The Journal of Aging and*
195 *Mental Health*. 1982;1(1):37–43.
- 196 47 Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: A preliminary report.
197 *Journal of Psychiatric Research*. 1982 Jan;17(1):37–49.
- 198 48 Verstraten P, Van Eekelen C. Handleiding voor de GIP: Gedragsobservatieschaal voor de Intramurale Psychogeriatric. Deventer: Van Loghum Slaterus; 1987.
- 199 49 Cucinotta D, Ambrosoli L, Poli A, Martorelli M, Savorani GC, Anzivino F. Clinical assessment of mental decline in elderly people: A proposal for a new
200 quantitative index. *Aging Clin Exp Res*. 1995 Feb;7(1):29–34.
- 201 50 Kolitz BP, Vanderploeg RD, Curtiss G. Development of the Key Behaviors Change Inventory: a traumatic brain injury behavioral outcome assessment
202 instrument. *Arch Phys Med Rehabil*. 2003 Feb;84(2):277–84.
- 203 51 Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: comprehensive assessment of
204 psychopathology in dementia. *Neurology*. 1994 Dec;44(12):2308–14.
- 205 52 Fahn S, Elton RL, Members of the UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. *Recent Developments in Parkinson's Disease*
206 *Volume II*. Macmillan Healthcare Information; 1987; pp 153–63.

- 207 53 Lee S-H, Wen M-C, Chao C-C, Chen YJ, Yen C-F. Apathy in late-life depression among Taiwanese patients. *Int Psychogeriatr*. 2008 Apr;20(2):328–37.
- 208 54 Clarke DE, Reekum R van, Simard M, Streiner DL, Freedman M, Conn D. Apathy in dementia: an examination of the psychometric properties of the apathy
209 evaluation scale. *J Neuropsychiatry Clin Neurosci*. 2007;19(1):57–64.
- 210 55 Guercio BJ, Donovan NJ, Munro CE, Aghjayan SL, Wigman SE, Locascio JJ, et al. The Apathy Evaluation Scale: A Comparison of Subject, Informant, and
211 Clinician Report in Cognitively Normal Elderly and Mild Cognitive Impairment. *J Alzheimers Dis*. 2015;47(2):421–32.
- 212 56 Johansson M, Johansson P, Stomrud E, Hagell P, Hansson O. Psychometric testing of a Swedish version of the Apathy Evaluation Scale. *Nord J Psychiatry*.
213 2017 Aug;71(6):477–84.
- 214 57 Lueken U, Evens R, Balzer-Geldsetzer M, Baudrexel S, Dodel R, Gräber-Sultan S, et al. Psychometric properties of the apathy evaluation scale in patients
215 with Parkinson’s disease. *Int J Methods Psychiatr Res*. 2017;26(4). DOI: 10.1002/mpr.1564
- 216 58 Morita H, Kannari K. Reliability and validity assessment of an apathy scale for home-care patients with Parkinson’s disease: a structural equation modeling
217 analysis. *J Phys Ther Sci*. 2016 Jun;28(6):1724–7.
- 218 59 Kay DB, Kirsch-Darrow L, Zahodne LB, Okun MS, Bowers D. Dimensions of apathy in Parkinson’s disease: exploratory factor analysis of the apathy scale. *J*
219 *Parkinsons Dis*. 2012;2(2):161–6.
- 220 60 Pedersen KF, Alves G, Larsen JP, Tysnes O-B, Møller SG, Brønneck K. Psychometric properties of the Starkstein Apathy Scale in patients with early untreated
221 Parkinson disease. *Am J Geriatr Psychiatry*. 2012 Feb;20(2):142–8.
- 222 61 Lopez FV, Eglit GML, Schiehser DM, Pirogovsky-Turk E, Litvan I, Lessig S, et al. Factor Analysis of the Apathy Scale in Parkinson’s Disease. *Movement*
223 *Disorders Clinical Practice*. 2019;6(5):379–86.
- 224 62 Santangelo G, D’Iorio A, Piscopo F, Cuoco S, Longo K, Amboni M, et al. Assessment of apathy minimising the effect of motor dysfunctions in Parkinson’s
225 disease: a validation study of the dimensional apathy scale. *Qual Life Res*. 2017;26(9):2533–40.
- 226 63 Radakovic R, McGrory S, Chandran S, Swingler R, Pal S, Stephenson L, et al. The brief Dimensional Apathy Scale: A short clinical assessment of apathy. *Clin*
227 *Neuropsychol*. 2020 Feb;34(2):423–35.

- 228 64 Midden AJ, Mast BT. IRT Analysis of the Frontal Systems Behavior Scale: Identifying the Best Items for Use with Geriatric Patients in Primary Care. *Clinical Gerontologist*. 2017 Aug;40(4):295–306.
229
- 230 65 García-Ramos R, Villanueva Iza C, Catalán MJ, Reig-Ferrer A, Matías-Guío J. Validation of a Spanish Version of the Lille Apathy Rating Scale for Parkinson's Disease [Internet]. *The Scientific World Journal*. 2014 Oct;2014:e849834.
231
- 232 66 Gallo JL, Schmidt KS, Libon DJ. An itemized approach to assessing behavioral and psychological symptoms in dementia. *Am J Alzheimers Dis Other Demen*. 2009 May;24(2):163–8.
233
- 234 67 Ang Y-S, Lockwood PL, Kienast A, Plant O, Drew D, Slavkova E, et al. Differential impact of behavioral, social, and emotional apathy on Parkinson's disease. *Annals of Clinical and Translational Neurology*. 2018;5(10):1286–91.
235
- 236 68 Carvalho JO. Revision of the frontal systems behavior scale (FrSBe). *Doctoral Dissertations Available from Proquest*. 2011 Jan;1–107.
- 237 69 Davis TE. The Effectiveness of the Geriatric Depression Scale to Distinguish Apathy From Depression in Alzheimer's Disease and Related Dementias. 2008
- 238 70 Dujardin K, Sockeel P, Delliaux M, Destée A, Defebvre L. The Lille Apathy Rating Scale: validation of a caregiver-based version. *Mov Disord*. 2008
239 Apr;23(6):845–9.
- 240 71 Köller L, Knebel M, Haberstroh J, Krause K, Sahlender S, Jakob M, et al. Apathy in dementia care: An investigation in community-dwelling persons using the
241 German informant version of the Apathy Evaluation Scale. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*. 2016;29(3):129–39.
- 242 72 Leone E, Lafont V, Filleau C, Baudu C, Benoit M, Deudon A, et al. Validation de la version soignant de l'Inventaire Apathie. *Revue Geriatr*. 2008;33(9):777–
243 85.
- 244 73 Malakouti SK, Panaghi L, Foroughan M, Salehi M, Zandi T. Farsi version of the Neuropsychiatric Inventory: validity and reliability study among Iranian
245 elderly with dementia. *Int Psychogeriatr*. 2012 Feb;24(2):223–30.
- 246 74 Radakovic R, Starr JM, Abrahams S. A Novel Assessment and Profiling of Multidimensional Apathy in Alzheimer's Disease. *J Alzheimers Dis*. 2017;60(1):57–
247 67.
- 248 75 Radakovic R, Davenport R, Starr JM, Abrahams S. Apathy dimensions in Parkinson's disease. *Int J Geriatr Psychiatry*. 2018 Jan;33(1):151–8.

- 249 76 Radakovic R, Gray D, Dudley K, Mioshi E, Dick D, Melchiorre G, et al. Reliability and validity of the brief dimensional apathy scale. *Arch Clin Neuropsychol*.
250 2020 Jul;35(5):539–44.
- 251 77 Serrano-Dueñas M, Martínez-Martín P, Merchán T, Bravo R, Serrano M. Properties of the Apathy Scale (AS) for use on Parkinson’s patients. *Advances in*
252 *Parkinson’s Disease*. 2013 May;2(2):53–7.
- 253 78 Stankevich Y, Lueken U, Balzer-Geldsetzer M, Dodel R, Gräber-Sultan S, Berg D, et al. Psychometric Properties of an Abbreviated Version of the Apathy
254 *Evaluation Scale for Parkinson Disease (AES-12PD)*. *The American Journal of Geriatric Psychiatry*. 2018 Oct;26(10):1079–90.
- 255 79 Vilalta-Franch J, Lozano-Gallego M, Hernández-Ferrándiz M, Llinàs-Reglà J, López-Pousa S, López OL. [The Neuropsychiatric Inventory. Psychometric
256 properties of its adaptation into Spanish]. *Rev Neurol*. 1999 Jul;29(1):15–9.
- 257 80 Stella F, Andrade LP de, Garuffi M, Vital TM, Hernández SSS, Ruocco M, et al. Validation of the Brazilian version of the Apathy Inventory. *International*
258 *Journal of Geriatric Psychiatry*. 2013;28(9):979–86.
- 259 81 Shinagawa Y, Nakaaki S, Hongo J, Murata Y, Sato J, Matsui T, et al. Reliability and validity of the Japanese version of the Dysexecutive Questionnaire (DEX)
260 in Alzheimer’s disease: validation of a behavioral rating scale to assess dysexecutive symptoms in Japanese patients with Alzheimer’s disease. *Int J Geriatr*
261 *Psychiatry*. 2007 Oct;22(10):951–6.
- 262 82 Diesfeldt HFA. [Interpreting change scores of the Behavioural Rating Scale for Geriatric Inpatients (GIP)]. *Tijdschr Gerontol Geriatr*. 2013 Sep;44(4):166–74.
- 263 83 Fernández-Matarrubia M, Matías-Guiu JA, Moreno-Ramos T, Valles-Salgado M, Marcos-Dolado A, García-Ramos R, et al. Validation of the Lille’s Apathy
264 *Rating Scale in Very Mild to Moderate Dementia*. *The American Journal of Geriatric Psychiatry*. 2016 Jul;24(7):517–27.
- 265 84 Camozzato AL, Kochhann R, Simeoni C, Konrath CA, Franz AP, Carvalho A, et al. Reliability of the Brazilian Portuguese version of the Neuropsychiatric
266 *Inventory (NPI) for patients with Alzheimer’s disease and their caregivers*. *International Psychogeriatrics*. 2008 Apr;20(2):383–93.
- 267 85 Davidsdottir SR, Snaedal J, Karlsdottir G, Atladottir I, Hannesdottir K. Validation of the Icelandic version of the Neuropsychiatric Inventory with Caregiver
268 *Distress (NPI-D)*. *Nord J Psychiatry*. 2012 Feb;66(1):26–32.
- 269 86 Leung VP, Lam LC, Chiu HF, Cummings JL, Chen QL. Validation study of the Chinese version of the neuropsychiatric inventory (CNPI). *Int J Geriatr Psychiatry*.
270 2001 Aug;16(8):789–93.

- 271 87 de Medeiros K, Robert P, Gauthier S, Stella F, Politis A, Leoutsakos J, et al. The Neuropsychiatric Inventory-Clinician rating scale (NPI-C): reliability and
272 validity of a revised assessment of neuropsychiatric symptoms in dementia. *Int Psychogeriatr*. 2010 Sep;22(6):984–94.
- 273 88 Stella F, Forlenza OV, Laks J, de Andrade LP, Ljubetic Avendaño MA, Gasparetto Sé EV, et al. The Brazilian version of the Neuropsychiatric Inventory-
274 Clinician rating (NPI-C): Reliability and validity in dementia. *Int Psychogeriatr*. 2013 Sep;25(9):1503–11.
- 275 89 Valentino V, Iavarone A, Amboni M, Moschiano F, Picillo M, Petretta V, et al. Apathy in Parkinson’s disease: differences between caregiver’s report and self-
276 evaluation. *Funct Neurol*. 2018 Mar;33(1):31–5.
- 277 90 Guimarães HC, Fialho PPA, Carvalho VA, Santos EL dos, Caramelli P, Guimarães HC, et al. Brazilian caregiver version of the Apathy Scale. *Dementia &
278 Neuropsychologia*. 2009 Dec;3(4):321–6.
- 279 91 Wetmore JB, Arbelo JM, Catalán MJ, Valldeoriola F, Rodríguez-Blázquez C, Martínez-Martin P. Psychometric Properties of the Apathy Scale in Advanced
280 Parkinson’s Disease [Internet]. *Parkinson’s Disease*. 2019 Mar;2019:e1965394.
- 281 92 Bertens AS, Moonen JEF, de Waal MWM, Foster-Dingley JC, de Ruijter W, Gussekloo J, et al. Validity of the three apathy items of the Geriatric Depression
282 Scale (GDS-3A) in measuring apathy in older persons. *Int J Geriatr Psychiatry*. 2017;32(4):421–8.
- 283 93 Belanger HG, Brown LM, Crowell TA, Vanderploeg RD, Curtiss G. The Key Behaviors Change Inventory and executive functioning in an elderly clinic sample.
284 *Clin Neuropsychol*. 2002 Aug;16(3):251–7.
- 285 94 Choi SH, Na DL, Kwon HM, Yoon SJ, Jeong JH, Ha CK. The Korean version of the neuropsychiatric inventory: a scoring tool for neuropsychiatric disturbance
286 in dementia patients. *J Korean Med Sci*. 2000 Dec;15(6):609–15.
- 287 95 Politis AM, Mayer LS, Passa M, Maillis A, Lyketsos CG. Validity and reliability of the newly translated Hellenic Neuropsychiatric Inventory (H-NPI) applied to
288 Greek outpatients with Alzheimer’s disease: a study of disturbing behaviors among referrals to a memory clinic. *Int J Geriatr Psychiatry*. 2004
289 Mar;19(3):203–8.
- 290 96 Kat MG, de Jonghe JFM, Aalten P, Kalisvaart CJ, Dröes RM, Verhey FRJ. [Neuropsychiatric symptoms of dementia: psychometric aspects of the Dutch
291 Neuropsychiatric Inventory (NPI)]. *Tijdschr Gerontol Geriatr*. 2002 Sep;33(4):150–5.

- 292 97 Weintraut R, Karádi K, Lucza T, Kovács M, Makkos A, Janszky J, et al. Lille Apathy Rating Scale and MDS-UPDRS for Screening Apathy in Parkinson's Disease.
293 J Parkinsons Dis. 2016;6(1):257–65.
- 294 98 Kirsch-Darrow L, Zahodne LB, Hass C, Mikos A, Okun MS, Fernandez HH, et al. How Cautious Should We Be When Assessing Apathy with the Unified
295 Parkinson's Disease Rating Scale? Mov Disord. 2009 Apr;24(5):684–8.
- 296 99 Pedersen KF, Larsen JP, Aarsland D. Validation of the Unified Parkinson's Disease Rating Scale (UPDRS) section I as a screening and diagnostic instrument
297 for apathy in patients with Parkinson's disease. Parkinsonism & Related Disorders. 2008 Apr;14(3):183–6.
- 298 100 Starkstein SE, Merello M. The Unified Parkinson's Disease Rating Scale: Validation study of the mentation, behavior, and mood section. Movement
299 Disorders. 2007;22(15):2156–61.
- 300 101 Cummings JL, Friedman JH, Garibaldi G, Jones M, Macfadden W, Marsh L, et al. Apathy in Neurodegenerative Diseases: Recommendations on the
301 Design of Clinical Trials. J Geriatr Psychiatry Neurol. 2015 Sep;28(3):159–73.
- 302 102 Marin RS. Apathy: a neuropsychiatric syndrome. JNP. 1991 Aug;3(3):243–54.
- 303 103 König A, Aalten P, Verhey F, Bensadoun G, Petit P-D, Robert PH, et al. A review of current information and communication technologies: can they be
304 used to assess apathy? Int J Geriatr Psychiatry. 2014 Apr;29(4):345–58.
- 305 104 Kolitz BP. Development of the Key Behaviors Change Inventory (KBCI), a TBI behavioral outcome assessment instrument. 2000 May
- 306 105 Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, et al. Movement Disorder Society-sponsored revision of the Unified
307 Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Mov Disord. 2007 Jan;22(1):41–7.
- 308 106 Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified
309 Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. 2008 Nov;23(15):2129–70.
- 310 107 GIP-28 Instruction Manual. [GIP-28 Handleiding]. [cited 2018 Oct 19]. Available from: [https://www.werkenindeouderengeneeskunde.nl/wp-](https://www.werkenindeouderengeneeskunde.nl/wp-content/uploads/2011/04/GIP-handleiding.pdf)
311 [content/uploads/2011/04/GIP-handleiding.pdf](https://www.werkenindeouderengeneeskunde.nl/wp-content/uploads/2011/04/GIP-handleiding.pdf)

