AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

© 2023 American Psychological Association
ISSN: 0894-4105

2024, Vol. 38, No. 1, 96–105
https://doi.org/10.1037/neu0000914

What's in a Score: A Longitudinal Investigation of Scores Based on Item Response Theory and Classical Test Theory for the Amsterdam Instrumental Activities of Daily Living Questionnaire in Cognitively Normal and Impaired Older Adults

Mark A. Dubbelman[1, 2], Merel C. Postema[1, 2], Roos J. Jutten[3], John E. Harrison[1, 2, 4, 5], Craig W. Ritchie[6], André Aleman[7], Frank Jan de Jong[8], Benjamin D. Schalet[9, 10], Caroline B. Terwee[9, 10], Wiesje M. van der Flier[1, 2, 9, 10], Philip Scheltens[1, 2], and Sietske A. M. Sikkes[1, 2, 11]

[1] Department of Neurology, Alzheimer Center Amsterdam, Vrije Universiteit Amsterdam, Amsterdam UMC
[2] Amsterdam Neuroscience, Neurodegeneration, Amsterdam, The Netherlands
[3] Department of Neurology, Massachusetts General Hospital, Harvard Medical School
[4] Metis Cognition Ltd, Wiltshire, United Kingdom
[5] Institute of Psychiatry, Psychology and Neuroscience, King's College London
[6] Centre for Dementia Prevention, University of Edinburgh
[7] Department of Neurosciences, University Medical Center Groningen, University of Groningen
[8] Department of Neurology, Erasmus Medical Center
[9] Epidemiology and Data Science, Vrije Universiteit Amsterdam, Amsterdam UMC
[10] Amsterdam Public Health, Methodology, Amsterdam, The Netherlands
[11] Department of Clinical, Neuro- and Developmental Psychology, Vrije Universiteit Amsterdam

Objective: We aimed to investigate whether item response theory (IRT)-based scoring allows for a more accurate, responsive, and less biased assessment of everyday functioning than traditional classical test theory (CTT)-based scoring, as measured with the Amsterdam Instrumental Activities of Daily Living Questionnaire. ***Method:*** In this longitudinal multicenter study including cognitively normal and impaired individuals, we examined IRT-based and CTT-based score distributions and differences between diagnostic groups using linear regressions, and investigated scale attenuation. We compared change over time between scoring methods using linear mixed models with random intercepts and slopes for time. ***Results:*** Two thousand two hundred ninety-four participants were included (66.6 ± 7.7 years, 54% female): n = 2,032 (89%) with normal cognition, n = 93 (4%) with subjective cognitive decline, n = 79 (3%) with mild cognitive impairment, and n = 91 (4%) with dementia. At baseline, IRT-based and CTT-based scores were highly correlated ($r = -0.92$). IRT-based scores showed less scale attenuation than CTT-based scores. In a subsample of n = 1,145 (62%) who were followed for a mean of 1.3 (SD = 0.6) years, IRT-based scores declined significantly among cognitively normal individuals (unstandardized coefficient [B] = −0.15, 95% confidence interval, 95% CI [−0.28, −0.03], effect size = −0.02), whereas CTT-based scores did not (B = 0.20, 95% CI [−0.02, 0.41], effect size = 0.02). In the other diagnostic groups, effect sizes of change over time were similar. ***Conclusions:*** IRT-based scores were less affected by scale attenuation than CTT-based scores. With regard to responsiveness, IRT-based scores showed more signal than CTT-based scores in early disease stages, highlighting the IRT-based scores' superior suitability for use in preclinical populations.

This article was published Online First September 7, 2023.
Mark A. Dubbelman https://orcid.org/0000-0002-5708-4925
Merel C. Postema https://orcid.org/0000-0002-1536-7062
Roos J. Jutten https://orcid.org/0000-0002-3712-4790
Caroline B. Terwee https://orcid.org/0000-0003-4570-2826
Wiesje M. van der Flier https://orcid.org/0000-0001-8766-6224
Sietske A. M. Sikkes https://orcid.org/0000-0002-0161-3656

This work was supported by grants from the Dutch Research Council (ZonMW) and the EU Joint Programme—Neurodegenerative Disease Research. In the past 2 years, John E. Harrison has received honoraria and paid consultancy from 23andMe, Abbvie, A2Q, AlzCure, Amgen, Anavex, Aptinyx, Astellas, AstraZeneca, Avraham, Axon, Axovant, Biogen, Boehringer Ingelheim, Bracket, Catenion, Cognition Therapeutics, CRF Health, Curasen, DeNDRoN, Enzymotec, Eisai, Eli Lilly, GfHEu, Heptares, Johnson and Johnson, Kaasa Health, Lysosome Therapeutics, Lundbeck, Merck, MyCognition, Mind Agilis, Neurocog, Neurim, Neuroscios, Neurotrack, Novartis, Nutricia, Pfizer, PriceSpective, Probiodrug, Regeneron, Rodin Therapeutics, Roche, Sanofi, Servier, Shire, Takeda, and vTv Therapeutics. Wiesje M. van der Flier's research programs have been funded by ZonMW, the Netherlands Organization of Scientific Research, Alzheimer Nederland, Cardiovasculair Onderzoek Nederland, Stichting Diorapthe, Gieskes-Strijbis fonds, Pasman stichting, Boehringer Ingelheim, Life-MI, AVID, Biogen MA, and Combinostics (all funding paid to her institution). Philip Scheltens has received consultancy fees (paid to his institution) from Alzheon, Brainstorm Cell, ImmunoBrain Checkpoint, and Green Valley. Within his university affiliation, he is principal investigator of studies with AC Immune, FUJI-film/Toyama, Ionis Pharmaceuticals, UCB, Novo Nordisk, and Vivoryon. He is also an employee of EQT Life Sciences (formerly LSP). Sietske A. M. Sikkes provided consultancy services for Boehringer Ingelheim, Takeda, and Nutricia; her work is supported

continued

Key Points
Question: Can we precisely detect small changes in everyday functioning in the context of Alzheimer's disease using an item response theory (IRT)-based scoring method? ***Findings:*** IRT-based scores of the Amsterdam Instrumental Activities of Daily Living Questionnaire showed a limited ceiling effect and were responsive to subtle decline in everyday functioning in the preclinical disease stage. ***Importance:*** With an increasing emphasis on preclinical disease stages, it is imperative to use outcome measures that have adequate psychometric properties in these populations and that are responsive to early changes. ***Next Steps:*** IRT-based computerized adaptive testing may reduce questionnaire completion times without losing measurement accuracy.

Keywords: dementia, Alzheimer's disease, instrumental activities of daily living, item response theory, outcome measure

Supplemental materials: https://doi.org/10.1037/neu0000914.supp

Dementia is the end stage of several neurodegenerative diseases, of which Alzheimer's disease (AD; Scheltens et al., 2021) is the most common. A core characteristic of dementia is impairment in performing cognitively complex activities, or so-called "instrumental activities of daily living" (IADL; Dubbelman, Jutten, et al., 2020; Marshall et al., 2017, 2020), such as managing paperwork and making appointments. The level of IADL functioning represents a clinically relevant outcome, even among those who have subjective cognitive complaints or mild cognitive impairment (Dubbelman, Jutten, et al., 2020; Marshall et al., 2012), which are considered prodromal disease stages preceding dementia.

The Amsterdam IADL Questionnaire (A-IADL-Q; Sikkes et al., 2012) was developed to assess difficulties in the performance of complex everyday activities due to cognitive decline. The A-IADL-Q has been extensively validated, showing good psychometric properties (Jutten et al., 2018; Jutten, Peeters, et al., 2017; Sikkes, Knol, et al., 2013; Sikkes, Pijnenburg, et al., 2013), in particular sensitivity to meaningful changes over time (Dubbelman, Verrijp, et al., 2022; Koster et al., 2015) and limited bias by culture, age, sex, or education level (Dubbelman, Verrijp, et al., 2020). The A-IADL-Q is scored using item response theory (IRT), which employs mathematical models to describe the relationship between a person's true ability on a construct that is not directly observable and the probability of the person giving a certain response to an individual item measuring that ability (Cella et al., 2010; Hays et al., 2000).

Theoretically, IRT holds several advantages of the more traditional classical test theory (CTT; Reise & Waller, 2009), which assumes that a person has a true ability that is measured with a certain degree of error. Importantly, it has been suggested that IRT-based scores may be less biased than CTT-based scores when estimating the change in a construct (Gorter et al., 2015, 2016; Jabrayilov et al., 2016), even potentially increasing responsiveness (Kosinski et al., 2003). Still, clinicians and regulatory agencies are reluctant to adopt IRT-based measures (Thomas, 2019), as CTT has the practical advantage that it is more straightforward to calculate the scores.

We aimed to investigate whether the advantages of IRT-based scoring allow for a more precise, less biased assessment of everyday functioning, as measured with the A-IADL-Q, in a predominantly cognitively normal sample. First, we aimed to examine score distributions and floor and ceiling effects between diagnostic groups. Second, we set out to analyze change over time in both scores. We hypothesized that the IRT-based scores (a) will discriminate well between people who are cognitively normal and those with subjective cognitive decline and (b) will show a stronger signal when assessing changes over time than CTT-based scores.

Materials and Method

Participants

We included participants from the European Prevention for Alzheimer's Disease Consortium Longitudinal Cohort Study

by grants from ZonMw and the European Union Joint Programme—Neurodegenerative Disease Research.

Sietske A. M. Sikkes and Philip Scheltens are codevelopers of the Amsterdam Instrumental Activities of Daily Living Questionnaire (A-IADL-Q). The A-IADL-Q is freely available for academic use and use by not-for-profit organizations. License fees from other uses of the A-IADL-Q are paid to their institution.

Mark A. Dubbelman played a lead role in formal analysis, visualization, writing–original draft, and writing–review and editing and a supporting role in conceptualization. Merel C. Postema played a supporting role in formal analysis and writing–review and editing. Roos J. Jutten played a supporting role in writing–review and editing and an equal role in data curation and project administration. John E. Harrison played a supporting role in methodology and writing–review and editing. Craig W. Ritchie played a supporting role in data curation and writing–review and editing and an equal role in funding acquisition. André Aleman played a supporting role in data curation and writing–review and editing. Frank Jan de Jong played a supporting role in data curation, funding acquisition, and writing–review and editing. Benjamin D. Schalet played a supporting role in writing–review and editing and an equal role in formal analysis. Caroline B. Terwee played a supporting role in conceptualization, formal analysis, and writing–review and editing. Wiesje M. van der Flier played a supporting role in data curation and writing–review and editing and an equal role in funding acquisition. Philip Scheltens played a supporting role in funding acquisition and writing–review and editing. Sietske A. M. Sikkes played a lead role in conceptualization and funding acquisition and a supporting role in data curation, formal analysis, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Sietske A. M. Sikkes, Department of Neurology, Alzheimer Center Amsterdam, Vrije Universiteit Amsterdam, Amsterdam UMC, Location VUmc, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands. Email: s.sikkes@amsterdamumc.nl

(Solomon et al., 2018), the Capturing Changes in Cognition study (Jutten, Harrison, et al., 2017), and the SCIENCe cohort (Slot et al., 2018). European Prevention for Alzheimer's Disease Consortium Longitudinal Cohort Study included participants without dementia from numerous research cohorts (e.g., brain health registries) and clinical or routine care cohorts (e.g., memory clinic) across Europe and was designed to reflect a trial-ready population. Participants were 50 years and older (Solomon et al., 2018). Participants were excluded if they met criteria for any type of dementia, were known carriers of autosomal dominant AD genes, had comorbid neurological or psychiatric disorders or cancer, had any contraindications for study procedures, had evidence of intracranial pathology, or were concurrently participating in a clinical trial. Participants in Capturing Changes in Cognition were recruited at multiple sites in The Netherlands and Scotland, were 65 years or older, and had diagnoses of subjective cognitive decline (SCD), mild cognitive impairment (MCI), or mild dementia at study inclusion (Jutten, Harrison, et al., 2017), as determined in a multidisciplinary consensus meeting using the criteria of the National Institute on Aging (Albert et al., 2011; McKhann et al., 2011). Exclusion criteria included comorbid neurological or psychiatric disorders, presence of depressive symptoms, current substance abuse, and concurrent participation in a clinical trial. Patients who visited the outpatient memory clinic of the Alzheimer Center Amsterdam were labeled as SCD in a multidisciplinary consensus meeting when they did not meet the diagnostic criteria of MCI (Petersen et al., 2014) or dementia (American Psychiatric Association, 2000) and had no major psychiatric disorder. They were included in the SCIENCe cohort (Slot et al., 2018). Exclusion criteria included comorbid neurological or psychiatric disorders, HIV, substance abuse, and the existence of a language barrier. In addition to the inclusion criteria of these studies, we selected individuals who had completed at least one A-IADL-Q, with item-level data available. When multiple assessments were available, we included repeated assessments, limiting follow-up to a maximum of 3 years from baseline. This study was performed in line with the principles of the Declaration of Helsinki. Approval for Capturing Changes in Cognition and SCIENCe was granted by the Ethics Committee of Vrije Universiteit University Medical Center. Approval for European Prevention for Alzheimer's Disease Consortium Longitudinal Cohort Study was approved by numerous institutional review boards across Europe. Informed consent to participate in the study was obtained from all participants.

Measures

A-IADL-Q

The A-IADL-Q is used to assess difficulties in the performance of various cognitively complex activities and is completed by a partner, adult child, or other observer (Sikkes et al., 2012). Items are rated on a scale from 0 = *no difficulty performing an activity* to 4 = *inability to perform the activity*. Missing values are introduced by design when the reason someone did not perform an activity was not related to cognitive decline (e.g., the person does not usually perform the activity, difficulties performing the activity were due to physical impairment). An example item with scoring and missing value assignment is displayed in Figure 1. IRT-based *T* scores are calculated using a graded response model (Samejima, 1969) and have a mean of 50 and a standard deviation of 10, with higher scores representing higher levels of everyday functioning. IRT item parameters, which include information on the location of the item on the latent trait and discriminatory ability, have been published by Jutten, Peeters, et al. (2017) and earlier by Sikkes, Pijnenburg, et al. (2013). CTT-based average scores are the sum of all items divided by the number of nonmissing items, which is multiplied by 25 so scores range from 0 to 100. Higher scores represent more severe impairment in everyday functioning.

We additionally aligned CTT-based scores to the IRT-based *T* scale. Because of the designed missingness and use of an average CTT-based score, a direct scale alignment approach, such as the Lord-Wingersky algorithm (Cai, 2015; Thissen et al., 1995), was not possible. Therefore, we randomly generated $n = 150{,}000$ responses to cover the entire spectrum of the scale, ranging from *severe* to *little or no impairment* in everyday functioning. Missingness was imposed according to missingness observed in the real-life A-IADL-Q data. The mean IRT-based score for each unique CTT-based score was taken as the aligned score, and these were used to create an alignment table. This table thus shows IRT-based scores corresponding to CTT-based scores.

More information on scoring and the IRT model and the scoring alignment procedure can be found in the Supplemental Material.

Data Analysis

Baseline differences between the diagnostic groups in IRT-based and CTT-based A-IADL-Q total scores were tested using analysis of variance and chi-squared tests, as appropriate, and *p* values were adjusted using Tukey's honest significant difference test. We also counted how often scale attenuation occurred at the extreme high ability (i.e., How many individuals had the lowest [on CTT-based scores] and highest scores [on IRT-based scores], representing no impairment?).

Next, we wanted to compare the ability of IRT-based and CTT-based A-IADL-Q scores to capture change over time. In the subsample of participants with longitudinal A-IADL-Q data, we ran two linear mixed models, one with IRT-based and another with CTT-based A-IADL-Q scores as the dependent variable. Time was the main independent variable, and we included random intercepts and random slopes for time. To examine differences between diagnostic groups, we also included a group by time interaction. All models were adjusted for age, sex, and education. To facilitate the comparison of change over time between scoring methods, we calculated effect sizes for a time by dividing the unstandardized time coefficient by the sum of the square roots of the variances of all intercepts and slopes, as well as the residual variance (Westfall et al., 2014).

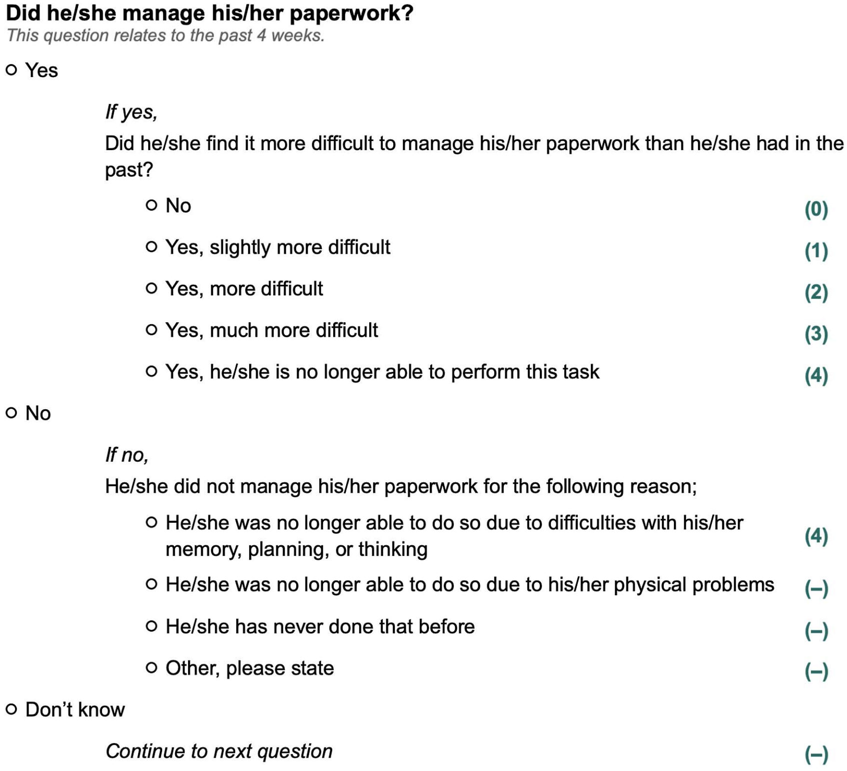
All analyses were performed in R Version 4.1.2 (R Core Team, 2022), using the "lme4" package Version 1.1-27 for the linear mixed models (Bates et al., 2015).

Transparency and Openness

This study was not preregistered. The data used in this article can be made available by the corresponding author upon reasonable request.

Figure 1

Example Item of the Amsterdam Instrumental Activities of Daily Living Questionnaire, With the Scoring Logic Displayed on the Right



Did he/she manage his/her paperwork?
This question relates to the past 4 weeks.

○ Yes

If yes,
Did he/she find it more difficult to manage his/her paperwork than he/she had in the past?

○ No (0)
○ Yes, slightly more difficult (1)
○ Yes, more difficult (2)
○ Yes, much more difficult (3)
○ Yes, he/she is no longer able to perform this task (4)

○ No

If no,
He/she did not manage his/her paperwork for the following reason;

○ He/she was no longer able to do so due to difficulties with his/her memory, planning, or thinking (4)
○ He/she was no longer able to do so due to his/her physical problems (–)
○ He/she has never done that before (–)
○ Other, please state (–)

○ Don't know

Continue to next question (–)

Note. (–) Denotes that the response is scored as missing. See the online article for the color version of this figure.

Results

We included 2,294 participants (66.6 ± 7.7 years old, 54% female, median education 15 years), most of whom (n = 2,031; 89%) were cognitively normal at inclusion. Of the remaining participants, 93 (35%) were diagnosed with SCD, 79 (30%) with MCI, and 91 (35%) with mild dementia. All diagnostic groups differed from each other in terms of age, with MCI participants being the oldest, followed by participants with dementia, normal cognition, and SCD (all adjusted $p < .01$). Sex distributions also differed between the groups ($p < .001$): Participants with normal cognition or dementia were more often female than others. Education differed significantly between participants with normal cognition and participants with SCD or dementia (both adjusted $p < .001$), but not between the other groups. Table 1 shows the baseline characteristics of the sample for the entire group, as well as each diagnostic group separately.

Baseline Differences

IRT-based and CTT-based scores correlated strongly (Pearson's $r = -0.92$, 95% confidence interval, 95% CI [−0.93, −0.91]). In cross-sectional comparisons, both IRT- and CTT-based A-IADL-Q scores were different between all groups (all adjusted $p < .001$; see Table 2). Based on the cross-sectional similarities, we created a crosswalk table to align IRT-based and CTT-based scores, which can be found in the Supplemental Material.

Figure 2 shows the baseline distributions of both scores for the different diagnostic groups. Scale attenuation affected the CTT-based scores of a total of 1,622 individuals (70.7%), who had a score of 0 indicating no impairment. Scale attenuation of the IRT-based scores occurred less often: Only 54 individuals (2.4%) had an IRT-based score of 70.0, which was the highest score reached in our sample, indicating no impairment. Further, while there were individuals in all diagnostic groups with a CTT-based score at the floor, only cognitively normal individuals reached the ceiling of IRT-based scores.

Change Over Time

A total of 1,415 individuals (61.7%) had longitudinal data available, with a median of two visits (interquartile range 2–3) per person and a mean of 1.35 ± 0.63 years of follow-up. Linear mixed models showed that both scoring techniques showed change over time in everyday functioning in the whole sample. Table 3 shows the unstandardized coefficients along with effect sizes for IRT- and CTT-based scores in the total sample and in the different diagnostic groups. IRT-based scores deteriorated modestly but significantly over time in cognitively normal older adults ($B = -0.15$, 95% CI [−0.28, −0.03]). Although CTT-based scores changed in the expected direction, this change did not reach significance ($B = 0.20$, 95% CI [−0.02, 0.41]). Both IRT- and CTT-based scores improved

Table 1
Baseline Characteristics

Characteristic	All	Normal cognition	Subjective cognitive decline	Mild cognitive impairment	Mild dementia
N	2,294	2,031 (88.5)	93 (4.1)	79 (3.4)	91 (4.0)
Age in years	66.6 ± 7.7	66.2 ± 7.5	63.5 ± 7.5	73.7 ± 7.9	70.9 ± 8.8
Female sex, *N* (%)	1,244 (54.2)	1,138 (56.0)	36 (37.9)	29 (36.7)	41 (45.1)
Education years, *Mdn* (IQR)	15 (12–17)	15 (12–17)	13 (10–17)	14 (12–16)	13 (10–16)
A-IADL-Q scores					
IRT, *M* ± *SD*	65.7 ± 6.9	67.3 ± 4.1	59.8 ± 8.8	54.3 ± 7.4	44.9 ± 8.6
IRT, range	29.8–70.0	42.1–70.0	35.3–69.8	31.6–69.7	29.8–68.6
IRT, *N* (%) at ceiling[a]	54 (2.4)	54 (2.7)	0 (0.0)	0 (0.0)	0 (0.0)
CTT, *M* ± *SD*	3.9 ± 11.8	1.3 ± 4.8	10.3 ± 15.7	17.2 ± 16.3	43.2 ± 25.5
CTT, range	0–87	0–53.3	0–71.5	0–78.3	0–87
CTT, *n* (%) at floor	1,622 (70.7)	1,589 (78.2)	27 (29.0)	5 (6.3)	1 (1.1)

Note. A-IADL-Q = Amsterdam Instrumental Activities of Daily Living Questionnaire; CTT = classical test theory; IRT = item response theory; IQR = interquartile range.
[a] The ceiling for the IRT-based scores was determined as 70.0: The highest scores achieved rounded to one decimal.

in SCD participants ($B = 1.33$, 95% CI [0.78, 1.89] and $B = -1.83$, 95% CI [−2.82, −0.83], respectively). In MCI and dementia, both IRT- and CTT-based scores deteriorated (see Table 3). Effect sizes were similar between the scoring methods in the whole group but were larger for CTT-based scores in more advanced disease stages. Figure 3 shows the change in the different scoring techniques over time, per diagnosis.

Discussion

In this study, we examined baseline and longitudinal differences between IRT-based and CTT-based scores of the A-IADL-Q. We found that they had largely similar distributions, but that IRT-based scores had much less of a ceiling effect than CTT-based scores, particularly in cognitively normal participants. In longitudinal analyses, effect sizes of change over time in both scores were comparable. IRT-based scores are less prone to ceiling effects than CTT-based scores, which is most evident in individuals with little impairment in daily functioning, where scores are more normally distributed and are responsive to small changes over time.

Table 2
Baseline Score Contrasts Between Diagnostic Groups for IRT-Based and CTT-Based Scores

Contrast	IRT	CTT
NC versus		
SCD	7.58 [6.59, 8.57]	−8.94 [−10.61, −7.27]
MCI	13.03 [11.96, 14.10]	−15.93 [−17.73, −14.13]
Dementia	22.46 [21.46, 23.46]	−41.89 [−43.57, −40.21]
SCD versus		
NC	−7.58 [−8.57, −6.59]	8.94 [7.27, 10.61]
MCI	5.46 [4.04, 6.88]	−6.99 [−9.39, −4.59]
Dementia	14.89 [13.52, 16.26]	−32.95 [−35.27, −30.63]
MCI versus		
NC	−13.03 [−14.10, −11.96]	15.93 [14.13, 17.73]
SCD	−5.46 [−6.88, −4.04]	6.99 [4.59, 9.39]
Dementia	9.43 [8.00, 10.86]	−25.96 [−28.38, −23.54]
Dementia versus		
NC	−22.46 [−23.46, −21.46]	41.89 [40.21, 43.57]
SCD	−14.89 [−16.26, −13.52]	32.95 [30.63, 35.27]
MCI	−9.43 [−10.86, −8.00]	25.96 [23.54, 28.38]

Note. IRT-based and CTT-based scores are not on the same scales and are mirrored to one another. CTT = classical test theory; IRT = item response theory; NC = normal cognition; MCI = mild cognitive impairment; SCD = subjective cognitive decline.

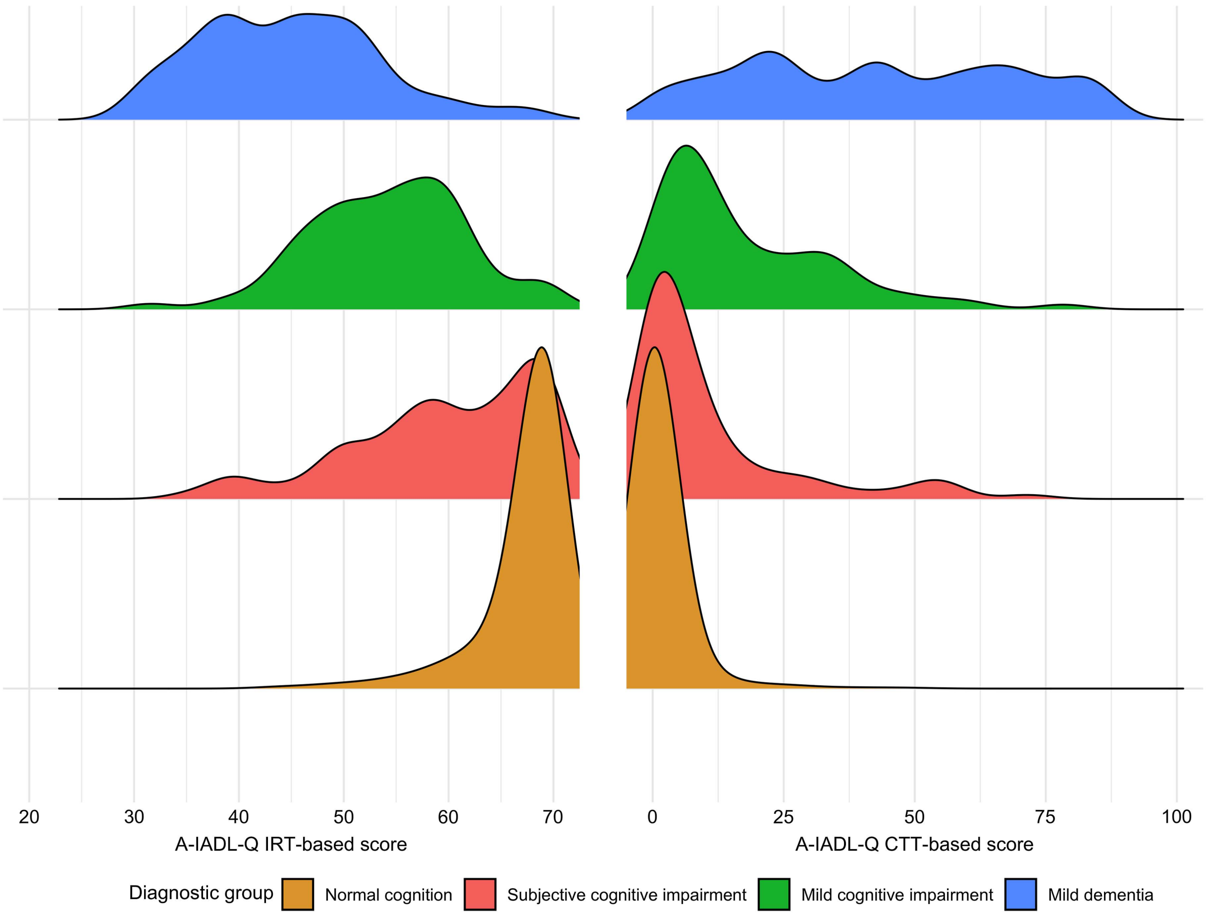
The IRT model sets apart the A-IADL-Q from many other functional measures, which are typically scored using CTT-based methods. Owing to the IRT parameters, previous studies of the A-IADL-Q have provided extensive validation, including good content and construct validity, diagnostic accuracy, and responsiveness (Koster et al., 2015; Sikkes, Knol, et al., 2013; Sikkes, Pijnenburg, et al., 2013; Verrijp et al., 2022). The A-IADL-Q is used internationally and has been culturally adapted; no systematic response bias has been found between countries (Bruderer-Hofstetter et al., 2020; Dubbelman, Verrijp, et al., 2020; Facal et al., 2018), and we determined thresholds for clinically meaningful changes (Dubbelman, Verrijp, et al., 2022). Much of this work would not have been possible without the IRT model. For practical reasons, in clinical practice, CTT-based scores are often used, as they can be calculated more easily, especially when the questionnaire is administered on a article. The question remained whether the IRT-based scoring method should be used in clinical practice as well, due to its hypothesized advantages.

Here, we showed that IRT- and CTT-based scores correlated almost perfectly, which is a well-established finding for other outcome measures (Reise & Waller, 2009). CTT-based scores showed substantial scale attenuation, with more than two thirds of all participants scoring at the extreme of the scale indicating no impairment. Similar effects have also been shown in other IADL instruments (Sikkes & Rotrou, 2014). Scale attenuation represents a lack of variance and poses a threat to the validity of analyses. For IRT-based scores, only two-and-a-half percent of participants scored at the extreme end of the scale indicating no impairment. As such, IRT-based scores are favored, especially in populations where the extremes of the scale are more frequently endorsed (i.e., in people who have no to very mild problems in everyday functioning).

There is a body of evidence suggesting that IRT-based scores outperform CTT-based scores in longitudinal analyses in terms of consistency of findings (Gorter et al., 2015, 2016; Jabrayilov et al., 2016), especially in small samples and in the face of missing data

Figure 2

Baseline Distributions of A-IADL-Q IRT-Based (Left) and CTT-Based (Right) Scores, Stratified by Diagnostic Group



Note. A-IADL-Q = Amsterdam Instrumental Activities of Daily Living Questionnaire; CTT = classical test theory; IRT = item response theory. See the online article for the color version of this figure.

(Wang & Reeve, 2021). As such, we had expected to find that IRT-based scores would show more evident change over time than CTT-based scores; however, the differences between the two methods in the amount of longitudinal change were minimal. Still, IRT-based scores showed a significant, albeit small, decline over time among cognitively normal individuals, whereas the change in CTT-based scores was not significant. The absence of scale attenuation in IRT-based scores may have contributed to this, as there was more variation in the IRT-based scores than in the CTT-based scores. This suggests that IRT-based scores might be

Table 3

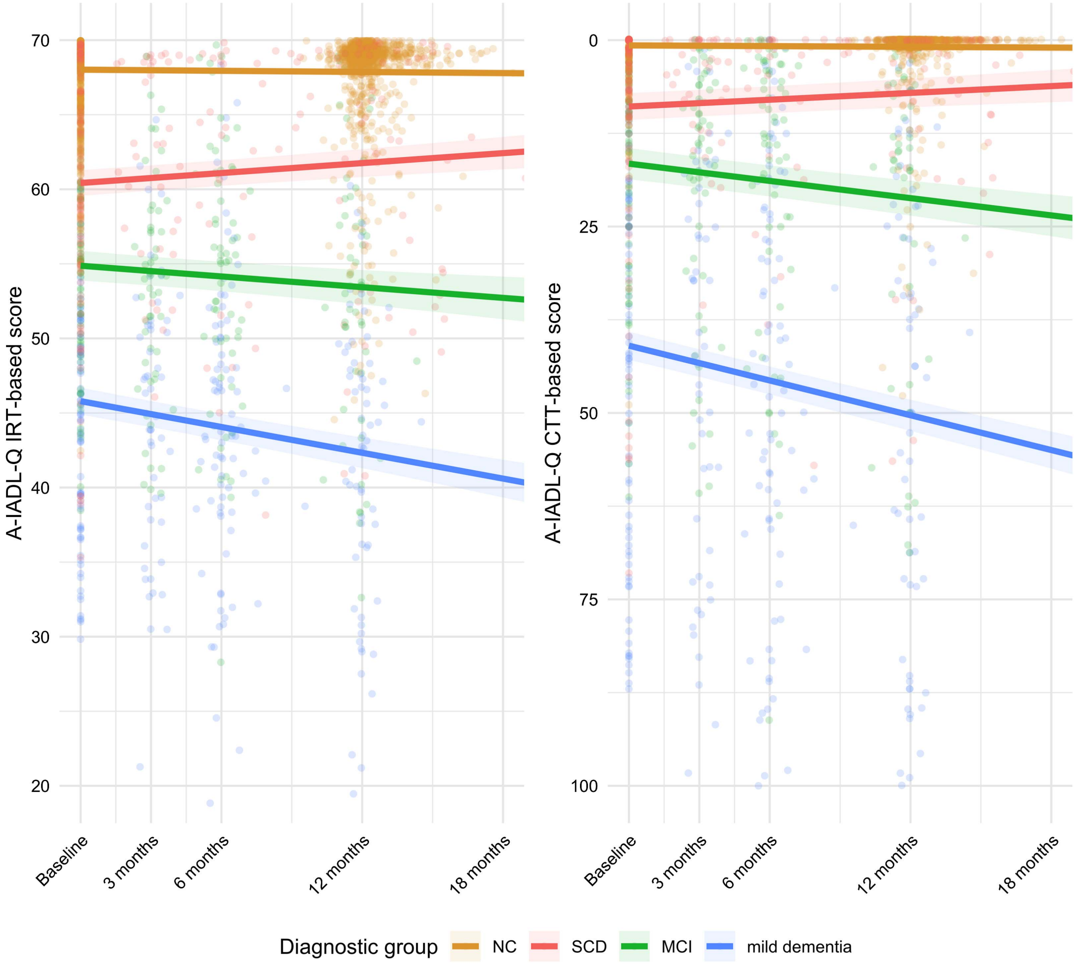
Estimated Yearly Change (Slopes) in IRT-Based and CTT-Based Scores in the Total Sample and in Different Groups

	IRT		CTT	
Scoring technique	Coefficient	Effect size	Coefficient	Effect size
Total group	−0.26 [−0.38, −0.13]	−0.24	0.68 [0.43, 0.93]	0.22
NC	−0.15 [−0.28, −0.03]	−0.02	0.20 [−0.02, 0.41]	0.02
SCD	1.33 [0.78, 1.89]	0.21	−1.83 [−2.82, −0.83]	−0.15
MCI	−1.44 [−2.29, −0.58]	−0.23	4.62 [3.09, 6.16]	0.37
Dementia	−3.46 [−4.21, −2.71]	−0.55	9.33 [7.97, 10.69]	0.74

Note. Time coefficients from linear mixed models are shown with the 95% confidence interval and are adjusted for baseline age, sex, and education. IRT and CTT score scales are mirrored to one another. CTT = classical test theory; IRT = item response theory; NC = normal cognition; MCI = mild cognitive impairment; SCD = subjective cognitive decline.

Figure 3

Trajectories of A-IADL-Q Scores Over Time, by Diagnostic Group



Note. Left panel: IRT scores, right panel: CTT average scores. The *y*-axis in the right panel is inversed so both scoring methods are in the same direction (i.e., a downward pointing line indicates a decline in everyday functioning). A-IADL-Q = Amsterdam Instrumental Activities of Daily Living Questionnaire; CTT = classical test theory; IRT = item response theory; MCI = mild cognitive impairment; NC = normal cognition; SCD = subjective cognitive decline. See the online article for the color version of this figure.

more suitable than CTT-based scores for tracking subtle changes over time. Interestingly, in individuals with SCD, we observed changes in the opposite direction, where functioning seemed to improve over time. This improvement was observed in both IRT- and CTT-based scores. Improvements in functioning are not commonly reported, but it is possible that reassurance after the initial memory clinic visit may have alleviated some concerns in this group of patients. At the same time, it should be noted the change was quite small. In more advanced disease stages, IRT- and CTT-based scores were comparable. In the dementia stage, CTT-based scores even showed a slightly larger effect size.

Together, these findings seem to indicate that the IRT-based scoring method has a modest advantage over CTT-based scoring, both for investigating cross-sectional differences and for measuring changes over time. The benefits of the IRT-based scoring method are particularly evident in early disease stages: It seems IRT- and CTT-based scores are more interchangeable in later disease stages, whereas IRT-based scores seem to be marginally more precise in early stages of AD and related disorders. Because CTT-based scores are easier to obtain, they could continue to be the preferred scoring method in clinical settings where there may not always be time or resources to run the IRT scoring algorithm. At the same time, we argue for the use of IRT-based scores in (secondary) prevention trials and research targeting early disease stages, as they appear to provide a more fine-grained assessment of IADL functioning.

Our study highlights methodological complexity when computing CTT to IRT crosswalk tables. There are various methods for linking CTT to IRT-based scores (Schalet et al., 2021), one of which is the Lord-Wingersky algorithm (Thissen et al., 1995), used in the PROsetta Stone project (Choi et al., 2021). The Lord-Wingersky algorithm relies on the IRT scoring parameters and was designed to link different instruments onto a single scale but may also be used to align different scoring techniques of the same instrument on the same scale. The latter works only with sum scores, but because of the design of the A-IADL-Q, in which responses that are not relevant for measuring the underlying construct of everyday functioning are considered missing, CTT-based scores are an average score. Hence, we could not use this algorithm to align the CTT and IRT-based scores. Therefore, we opted to use simulations instead, with the drawback that we could not simulate all possible item response combinations due to computational constraints with the enormous number of possible response patterns. An important advantage of aligning the two scales is that previously published IRT score cutoffs can be applied to CTT-based scores as well. Hence, we recommend users of the CTT-based scores to use the crosswalk tables to convert their scores into IRT-based *T* scores. Such crosswalk tables can be integrated into electronic health records or case report forms so that CTT-based scores can be converted automatically into and displayed as IRT-based *T* scores. In the transition from CTT-based to IRT-based scoring, with the linkage provided here, we eliminate the need for separate cutoffs, which thus far have only been made to apply to the IRT scale (Dubbelman, Terwee, et al., 2022; Dubbelman, Verrijp, et al., 2022; Sikkes, Pijnenburg, et al., 2013).

Constraints on Generality

Our sample was recruited in Western European countries, and we do not know the ethnoracial composition of our sample. Our sample was relatively highly educated. We previously showed that there was no meaningful systematic bias in the reporting of impairment in everyday functioning in individuals from various European countries and the United States nor in individuals who received more or fewer years of formal education (Dubbelman, Verrijp, et al., 2020). Therefore, we expect similar results in samples from the United States and among those who are less highly educated. It should be noted, however, that it is currently unknown how these findings generalize to the global population, as this has not yet been investigated. Further, we used data from multiple cohorts that each employed differing inclusion criteria. While this allowed us to obtain data from individuals covering a wider spectrum of cognitive statuses ranging from normal cognition to dementia, it is possible that cohort-specific inclusion criteria other than cognitive status might have influenced the clinical group comparisons. Finally, we were unable to account for the potential confounding effect of patient characteristics such as mood or comorbidity due to a lack of available data. We have previously observed a negligible influence of mood on everyday functioning (Dubbelman, Verrijp, et al., 2020; Sikkes, Knol, et al., 2013), hence, the confounding effect in the present analyses is likely limited.

Strengths and Future Directions

A strength of this study was the inclusion of a large sample of patients from various European sites who were followed over time, spanning the disease spectrum from cognitively normal to mild dementia and thus representing a wide array of different clinical presentations. Future studies with the A-IADL-Q may capitalize on other opportunities provided by IRT-based scoring, including computerized adaptive testing that may substantially reduce the burden of completing questionnaires. Based on item parameters that provide information on each item's difficulty and discriminatory ability, a precise total score can be obtained with far fewer—often less than 10—questions.

Conclusion

In conclusion, IRT-based scores for the A-IADL-Q have advantages including the lack of a ceiling effect, the possibility of computerized adaptive testing, and slightly superior responsiveness in early disease stages. As such, IRT-based scoring should be performed whenever possible to ensure that data can be analyzed optimally.

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev. ed.).

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bruderer-Hofstetter, M., Dubbelman, M. A., Meichtry, A., Koehn, F., Münzer, T., Jutten, R. J., Scheltens, P., Sikkes, S. A. M., & Niedermann, K. (2020). Cross-cultural adaptation and validation of the Amsterdam Instrumental Activities of Daily Living questionnaire short version German for Switzerland. *Health and Quality of Life Outcomes*, *18*(1), Article 323. https://doi.org/10.1186/s12955-020-01576-w

Cai, L. (2015). Lord-Wingersky Algorithm Version 2.0 for Hierarchical Item Factor Models with Applications in Test Scoring, Scale Alignment, and Model Fit Testing. *Psychometrika*, *80*(2), 535–559. https://doi.org/10.1007/s11336-014-9411-3

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., … the PROMIS Cooperative Group. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Choi, S. W., Lim, S., Schalet, B. D., Kaat, A. J., & Cella, D. (2021). PROsetta: An *R* package for linking patient-reported outcome measures. *Applied Psychological Measurement*, *45*(5), 386–388. https://doi.org/10.1177/01466216211013106

Dubbelman, M. A., Jutten, R. J., Tomaszewski Farias, S. E., Amariglio, R. E., Buckley, R. F., Visser, P. J., Rentz, D. M., Johnson, K. A., Properzi, M. J., Schultz, A., Donovan, N., Gatchell, J. R., Teunissen, C. E., Van Berckel, B. N. M., Van der Flier, W. M., Sperling, R. A., Papp, K. V., Scheltens, P., Marshall, G. A., … the Alzheimer Disease Neuroimaging Initiative, National Alzheimer's Coordinating Center, the Harvard Aging Brain Study, the Alzheimer Dementia Cohort. (2020). Decline in cognitively complex everyday activities accelerates along the Alzheimer's disease continuum. *Alzheimer's Research & Therapy*, *12*(1), Article 138. https://doi.org/10.1186/s13195-020-00706-2

Dubbelman, M. A., Terwee, C. B., Verrijp, M., Visser, L. N. C., Scheltens, P., & Sikkes, S. A. M. (2022). Giving meaning to the scores of the Amsterdam instrumental activities of daily living questionnaire:

A qualitative study. *Health and Quality of Life Outcomes*, *20*(1), Article 47. https://doi.org/10.1186/s12955-022-01958-2

Dubbelman, M. A., Verrijp, M., Facal, D., Sánchez-Benavides, G., Brown, L. J. E., van der Flier, W. M., Jokinen, H., Lee, A., Leroi, I., Lojo-Seoane, C., Milošević, V., Molinuevo, J. L., Pereiro Rozas, A. X., Ritchie, C., Salloway, S., Stringer, G., Zygouris, S., Dubois, B., Epelbaum, S., … Sikkes, S. A. M. (2020). The influence of diversity on the measurement of functional impairment: An international validation of the Amsterdam IADL Questionnaire in eight countries. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *12*(1), Article e12021. https://doi.org/10.1002/dad2.12021

Dubbelman, M. A., Verrijp, M., Terwee, C. B., Jutten, R. J., Postema, M. C., Barkhof, F., Berckel, B. N. M., Gillissen, F., Teeuwen, V., Teunissen, C., van de Flier, W. M., Scheltens, P., & Sikkes, S. A. M. (2022). Pursuing clinical meaningfulness: Determining the minimal important change of everyday functioning in dementia. *Neurology*, *99*(9), e954–e964. https://doi.org/10.1212/WNL.0000000000200781

Facal, D., Carabias, M. A. R., Pereiro, A. X., Lojo-Seoane, C., Campos-Magdaleno, M., Jutten, R. J., Sikkes, S. A. M., & Juncos-Rabadán, O. (2018). Assessing everyday activities across the dementia spectrum with the Amsterdam IADL Questionnaire. *Current Alzheimer Research*, *15*(13), 1261–1266. https://doi.org/10.2174/1567205015666180925113411

Gorter, R., Fox, J. P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology*, *79*, 140–149. https://doi.org/10.1016/j.jclinepi.2016.06.011

Gorter, R., Fox, J. P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, *15*(1), Article 55. https://doi.org/10.1186/s12874-015-0050-x

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(Suppl. 9), II-28–II-42. https://doi.org/10.1097/00005650-200009002-00007

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Jutten, R. J., Dicks, E., Vermaat, L., Barkhof, F., Scheltens, P., Tijms, B. M., & Sikkes, S. A. M. (2018). Impairment in complex activities of daily living is related to neurodegeneration in Alzheimer's disease-specific regions. *Neurobiol Aging*, *75*, 109–116. https://doi.org/10.1016/j.neurobiolaging.2018.11.018

Jutten, R. J., Harrison, J., de Jong, F. J., Aleman, A., Ritchie, C. W., Scheltens, P., & Sikkes, S. A. M. (2017). A composite measure of cognitive and functional progression in Alzheimer's disease: Design of the Capturing Changes in Cognition study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *3*(1), 130–138. https://doi.org/10.1016/j.trci.2017.01.004

Jutten, R. J., Peeters, C. F. W., Leijdesdorff, S. M. J., Visser, P. J., Maier, A. B., Terwee, C. B., Scheltens, P., & Sikkes, S. A. M. (2017). Detecting functional decline from normal aging to dementia: Development and validation of a short version of the Amsterdam IADL Questionnaire. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *8*(1), 26–35. https://doi.org/10.1016/j.dadm.2017.03.002

Kosinski, M., Bjorner, J. B., Ware, J. E., Jr., Batenhorst, A., & Cady, R. K. (2003). The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: A re-analysis of three clinical trials. *Quality of Life Research*, *12*(8), 903–912. https://doi.org/10.1023/A:1026111029376

Koster, N., Knol, D. L., Uitdehaag, B. M., Scheltens, P., & Sikkes, S. A. M. (2015). The sensitivity to change over time of the Amsterdam IADL Questionnaire©. *Alzheimer's & Dementia*, *11*(10), 1231–1240. https://doi.org/10.1016/j.jalz.2014.10.006

Marshall, G. A., Aghjayan, S. L., Dekhtyar, M., Locascio, J. J., Jethwani, K., Amariglio, R. E., Johnson, K. A., Sperling, R. A., & Rentz, D. M. (2017). Activities of daily living measured by the Harvard Automated Phone Task track with cognitive decline over time in non-demented elderly. *The Journal of Prevention of Alzheimer's Disease*, *4*(2), 81–86. https://doi.org/10.14283/jpad.2017.10

Marshall, G. A., Amariglio, R. E., Sperling, R. A., & Rentz, D. M. (2012). Activities of daily living: Where do they fit in the diagnosis of Alzheimer's disease? *Neurodegenerative Disease Management*, *2*(5), 483–491. https://doi.org/10.2217/nmt.12.55

Marshall, G. A., Sikkes, S. A. M., Amariglio, R. E., Gatchel, J. R., Rentz, D. M., Johnson, K. A., Langford, O., Sun, C. K., Donohue, M. C., Raman, R., Aisen, P. S., Sperling, R. A., Galasko, D. R. (2020). Instrumental activities of daily living, amyloid, and cognition in cognitively normal older adults screening for the A4 Study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *12*(1), Article e12118. https://doi.org/10.1002/dad2.12118

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Jr., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263–269. https://doi.org/10.1016/j.jalz.2011.03.005

Petersen, R. C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., & Fratiglioni, L. (2014). Mild cognitive impairment: A concept in evolution. *Journal of Internal Medicine*, *275*(3), 214–228. https://doi.org/10.1111/joim.12190

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.1.3). https://www.R-project.org/

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*(1), 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/BF03372160

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*, *86*(3), 717–746. https://doi.org/10.1007/s11336-021-09776-z

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. *Lancet*, *397*(10284), 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4

Sikkes, S. A. M., de Lange-de Klerk, E. S., Pijnenburg, Y. A. L., Gillissen, F., Romkes, R., Knol, D. L., Uitdehaag, B. M., & Scheltens, P. (2012). A new informant-based questionnaire for instrumental activities of daily living in dementia. *Alzheimer's & Dementia*, *8*(6), 536–543. https://doi.org/10.1016/j.jalz.2011.08.006

Sikkes, S. A. M., Knol, D. L., Pijnenburg, Y. A. L., de Lange-de Klerk, E. S., Uitdehaag, B. M., & Scheltens, P. (2013). Validation of the Amsterdam IADL Questionnaire©, a new tool to measure instrumental activities of daily living in dementia. *Neuroepidemiology*, *41*(1), 35–41. https://doi.org/10.1159/000346277

Sikkes, S. A. M., Pijnenburg, Y. A. L., Knol, D. L., de Lange-de Klerk, E. S., Scheltens, P., & Uitdehaag, B. M. (2013). Assessment of instrumental activities of daily living in dementia: Diagnostic value of the Amsterdam Instrumental Activities of Daily Living Questionnaire. *Journal of Geriatric Psychiatry and Neurology*, *26*(4), 244–250. https://doi.org/10.1177/0891988713509139

Sikkes, S. A. M., & Rotrou, J. (2014). A qualitative review of instrumental activities of daily living in dementia: What's cooking? *Neurodegenerative Disease Management*, *4*(5), 393–400. https://doi.org/10.2217/nmt.14.24

Slot, R. E. R., Verfaillie, S. C. J., Overbeek, J. M., Timmers, T., Wesselman, L. M. P., Teunissen, C. E., Dols, A., Bouwman, F. H., Prins, N. D., Barkhof, F., Lammertsma, A. A., Van Berckel, B. N. M., Scheltens, P., Sikkes, S. A. M., & Van der Flier, W. M. (2018). Subjective Cognitive Impairment Cohort (SCIENCe): Study design and first results. *Alzheimer's Research & Therapy*, *10*(1), Article 76. https://doi.org/10.1186/s13195-018-0390-y

Solomon, A., Kivipelto, M., Molinuevo, J. L., Tom, B., Ritchie, C. W., & the EPAD Consortium. (2018). European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): Study protocol. *BMJ Open*, *8*(12), Article e021017. https://doi.org/10.1136/bmjopen-2017-021017

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39–49. https://doi.org/10.1177/014662169501900105

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Verrijp, M., Dubbelman, M. A., Visser, L. N. C., Jutten, R. J., Nijhuis, E. W., Zwan, M. D., van Hout, H. P. J., Scheltens, P., van der Flier, W. M., & Sikkes, S. A. M. (2022). Everyday functioning in a community-based volunteer population: Differences between participant- and study partner-report. *Frontiers in Aging Neuroscience*, *13*, Article 761932. https://doi.org/10.3389/fnagi.2021.761932

Wang, M., & Reeve, B. B. (2021). Evaluations of the sum-score-based and item response theory-based tests of group mean differences under various simulation conditions. *Statistical Methods in Medical Research*, *30*(12), 2604–2618. https://doi.org/10.1177/09622802211043263

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Received December 2, 2022
Revision received April 26, 2023
Accepted May 1, 2023 ■