Research Article

Development of robust normative data for the neuropsychological assessment of Greek older adults

Xanthi Arampatzi[1,2], Eleni S. Margioti[1,2], Lambros Messinis[1], Mary Yannakoulia[3], Georgios Hadjigeorgiou[4], Efthimios Dardiotis[5], Paraskevi Sakka[2], Nikolaos Scarmeas[6,7] and Mary H. Kosmidis[1] (iD)

[1]Lab of Cognitive Neuroscience, School of Psychology, Aristotle University of Thessaloniki, Thessaloniki, Greece, [2]Athens Alzheimer's Association, Athens, Greece, [3]Department of Nutrition and Dietetics, Harokopio University, Athens, Greece, [4]School of Medicine, University of Cyprus, Nicosia, Cyprus, [5]School of Medicine, University of Thessaly, Larissa, Greece, [6]1st Department of Neurology, Aiginition Hospital, National and Kapodistrian University of Athens Medical School, Athens, Greece and [7]Department of Neurology, Taub Institute for Research in Alzheimer's Disease and the Aging Brain, The Gertrude H. Sergievsky Center, Columbia University, New York, USA

Abstract

Objective: Normative data for older adults may be tainted by inadvertent inclusion of undiagnosed individuals at the very early stage of a neurodegenerative process. To avoid this pitfall, we developed norms for a cohort of older adults without MCI/dementia at 3-year follow-up. **Methods:** A randomly selected sample of 1041 community-dwelling individuals (age ≥ 65) received a full neurological and neuropsychological examination on two occasions [mean interval = 3.1 (SD = 0.9) years]. **Results:** Of these, 492 participants (Group 1; 65–87 years old) were without dementia on both evaluations (CDR=0 and MMSE ≥ 26); their baseline data were used for norms development. Group 2 ($n = 202$) met the aforementioned criteria only at baseline, but not at follow-up. Multiple linear regressions included demographic predictors for regression-based normative formulae and raw test scores as dependent variables for each test variable separately. Standardized scaled scores and stratified discrete norms were also calculated. Group 2 performed worse than Group 1 on most tests (p-values < .001–.021). Education was associated with all test scores, age with most, and sex effects were consistent with the literature. **Conclusions:** We provide a model for developing sound normative data for widely used neuropsychological tests among older adults, untainted by potential early, undiagnosed cognitive impairment, reporting regression-based, scaled, and discrete norms for use in clinical settings to identify cognitive decline in older adults. Additionally, our co-norming of a variety of tests may enable intra-individual comparisons for diagnostic purposes. The present work addresses the challenge of developing robust normative data for neuropsychological tests in older adults.

Keywords: aging; neuropsychological impairment; normative data; older adults; co-norming

Introduction

Normative data for neuropsychological tests are commonly used in clinical assessments to evaluate an individual's cognitive abilities relative to a healthy comparison group. When dealing with older adults, however, healthy comparison groups may inadvertently include individuals at the very early preclinical stages of a neurodegenerative disorder, such as Mild Cognitive Impairment (MCI) or dementia. Inclusion of such cases in a standardization sample may reduce normative standards and potentially lead to erroneous classification as cognitively healthy (McCaffrey & Westervelt, 1995; Oosterhuis et al., 2016; Salthouse & Tucker-Drob, 2008). Thus, sensitive measures and methods of detection are needed to ensure accurate early identification of such disorders.

Given the insidious nature of neurodegenerative disorders among older individuals, cross-sectional studies in this age group, by their nature, cannot supply adequate information about possible cases with early underlying pathology. Instead, longitudinal studies are needed to confirm, on follow-up, the non-neurological status of participants in the original normative databases. Thus, ideally, a normative study of older adults should include a longitudinal investigation to allow the exclusion *ex post facto* of any individuals from the initial assessment who later developed dementia to ensure that the normative data are based on healthy individuals. Few studies, to date, have developed such robust normative data (i.e., Alden et al., 2022; Kramer et al., 2020) based on samples of older adults who were clinically normal across several assessments, yielding norms that were more sensitive to early detection of memory decline than existing norms.

Another critical issue in developing normative data, which may influence diagnostic accuracy, relates to the norming procedure. Discrete norms are the most widely available and frequently used. Particular demographic characteristics (e.g., age,

educational level, sex) have been found to influence performance on most neuropsychological tests and they are routinely controlled for in discrete normative databases (Lezak et al., 2012; Mitrushina et al., 2005; Strauss et al., 2006). Traditional discrete norms tables typically provide data broken down by age and education, and sometimes sex (if this factor was found to influence performance on the particular test in question). Alternatively, others provide cutoff scores with or without corrective factors (e.g., add 1 or 2 points to an individual's score for low education or old age before comparing with normative data) (Duff & Ramezani, 2015; Heaton et al., 2004). Then, an individual's scores are compared with the appropriate normative group (i.e., people in the same age and education range) to determine whether his or her score deviates from that expected based on the performance of a healthy cohort with similar demographic characteristics.

Although quite popular, discrete norms have some limitations. One is that normative groups are typically divided based on age and educational level, which may sometimes be arbitrary. Another limitation is that performance designation can shift depending on which age band is used even though the raw score remains the same. For example, if an individual is 65 years old and is compared with a normative group between 65 and 80 years old, she or he may be at an advantage; yet, if this person's score were compared to the immediately preceding group (e.g., 50–65), her or his scores might indicate impairment. Furthermore, in some cases, published normative data classify older adults as a single age group with an age range of 60–90 years. Given the rapid decline in most cognitive abilities that comes with age, an individual's performance could either be overestimated or underestimated depending on where it is classified in the broad age range (overestimating impairment in the low range and underestimating it in the high age range). Finally, stratified discrete norms also have the disadvantage of requiring large sample sizes for reliable results due to statistical limitations (Duff et al., 2010; Oosterhuis et al., 2016).

An alternative approach is to derive continuous norms using multiple regression equations (Parmenter et al., 2010). This method, now commonly used in test development, is the Standard Regression-Based change score approach and has been used extensively in healthy as well as clinical samples (Brandt & Benedict, 2001; Martin et al., 2002). Regression-based normative formulae simultaneously correct for (via multivariate regression analyses) any demographic variables that may influence neuro-psychological performance (Temkin et al., 1999), and, thus, may be more sensitive than discrete norms in detecting late-life cognitive decline (Duff & Ramezani, 2015). Another benefit of developing regression-based or continuous norms is that they require smaller samples than are needed for discrete norms (Oosterhuis et al., 2016).

Therefore, our goal in undertaking the present study was initially to verify whether a group of individuals who demonstrably were not at an early stage of a neurodegenerative process would in fact yield stricter criteria than a group of older individuals who later developed MCI or dementia, and, thus, may have already been demonstrating subtle indications of decline. To our knowledge, this has only been done to date for single tests (i.e., Rey Auditory Verbal Learning Test in Alden et al., 2022; California Verbal Learning Test in Kramer et al., 2020). Yet clinical assessments typically rely on a battery of tests from which to come to potential diagnostic conclusions. Thus, we used a comprehensive test battery assessing simultaneously five major cognitive domains. If

confirmed, our next goal was to develop normative data based on the former, untainted group applying regression-based normative formulae, simultaneously correcting for all relevant demographic variables (age, level of education, and sex) and traditional discrete norms (stratified by the aforementioned demographic variables, if relevant), to compare the relative utility of each (based on patient examples) and provide clinicians with a choice for their clinical practice or research purposes.

Methods

Participants

Participants were recruited for the Hellenic Longitudinal Investigation of Aging and Diet (HELIAD), a longitudinal, epidemiologic, and population-based study in Greece, randomly selected from municipality rosters (city of Larissa and surrounding rural regions, as well as Marousi, a suburb of Athens). Procedures and study details have been described previously (Bougea et al., 2019; Dardiotis et al., 2014; Kosmidis et al., 2018; Vlachos et al., 2020). Briefly, of the randomly selected individuals 65 years of age and older, we were able to contact 3552, of whom 2004 (58% were women) agreed to participate. No exclusion criteria were applied besides age. About two-thirds of the total study sample had a low level of education (some primary education).

All participants provided written informed consent before their participation. The present research received approval by the ethics committees of the University of Thessaly and the National and Kapodistrian University of Athens and was conducted in accordance with the Helsinki Declaration.

Procedure

Cognitive status was evaluated on two occasions to eliminate the possibility of subsequent cognitive decline not yet detected at baseline evaluation. At baseline, neurologists conducted a standard physical and neurological examination, including a structured face-to-face interview regarding information about medical, neurological, psychiatric, and family history, as well as activities of daily living, social and leisure activities, and lifestyle patterns. The neurologists assigned a global summary score for each participant, namely, a Clinical Dementia Rating (CDR) score (Morris, 1993). Additionally, neuropsychologists administered a comprehensive neuropsychological test battery, which included tests of orientation, memory (verbal and non-verbal), language (naming, verbal fluency, comprehension, and repetition), attention/processing speed, executive functioning, visuospatial perception, and a gross estimate of intellectual abilities (vocabulary test).

Consensus diagnostic meetings were held to determine the potential presence of dementia (based on Diagnostic and Statistical Manual of Mental Disorders-IV-TR criteria; APA, 2000) and MCI (based on National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association criteria (McKhann et al., 1984) and International Working Group on MCI). Participants were reassessed after a mean time interval of approximately 3.1 (range: 1.3–7.3) years, repeating the baseline neurological and neuro-psychological assessment. After the follow-up examination, a consensus meeting was held to revise the participants' initial clinical diagnosis, if necessary.

Participants included in the present analyses met the following criteria: (1) they were considered not to have dementia at the baseline

diagnostic evaluation, based on the following two indices: a value of 0 on the CDR *and* a threshold value of ≥ 26 on the Mini-Mental State Examination (MMSE; Folstein et al., 1975), (2) they had returned for the follow-up evaluation and had CDR and MMSE ratings (regardless of their scores *per se*), and (3) they had completed the neuropsychological battery at least on the initial assessment. Illiterate individuals with no formal schooling (determined by self-report) were excluded from the present analyses and have been presented separately (Mandyla et al., 2022), as they pose not only a quantitatively but also a qualitatively, different group warranting specialized attention.

Neuropsychological assessment

For the purpose of the overall investigation, we compiled tests from multiple sources to provide a comprehensive assessment of cognitive functioning, memory, language, attention/processing speed, executive functioning and motor programing, visuospatial ability, and a gross estimate of general intelligence. The battery comprised the following tests (all those containing verbal items were Greek versions):

a. Medical College of Georgia Complex Figure Test (MCG; Lezak et al., 2004), consisting of a complex line drawing; variables of interest: correct number of items on copy and percentage of items copied correctly, immediate and delayed recall and percentage of items recalled on each condition based on the copy score, and recognition (created for the present study) conditions. We used version 1 (of the four versions available) during the first assessment and version 2 on the second assessment.

b. Greek Verbal Learning Test (GVLT; Vlahou et al., 2013), a list-learning test of semantically related items (16 in total, from 4 semantic categories); variables of interest: correct number of words on the first learning trial, the total of five learning trials, immediate and delayed recall (on both free and cued recall conditions), learning curve, and encoding, consolidation, retrieval deficit indices. We used test version A on the initial assessment and one of the three alternate forms (specifically, B) on the second assessment.

c. Verbal fluency: semantic (objects) and phonological verbal fluency (letter A) (Kosmidis et al., 2004); variables of interest: number of words, clusters, and switches produced for each condition. We used alternate forms in the two assessments.

d. Subtests of the Boston Diagnostic Aphasia Examination short form, specifically, the Boston Naming Test-short form (15-item), and selected items from the Complex Ideational Material Subtest, to assess verbal comprehension and repetition of three multisyllabic words and three short phrases, selected to be challenging to reproduce (Tsapkini et al., 2010); variables of interest: number correct on naming, comprehension, and repetition.

e. Trail Making Test (TMT), Parts A and B (Vlahou & Kosmidis, 2002); variables of interest: time to completion on each part and their ratio (TMT B:TMT A). For those who could not complete the test in 5 minutes (maximum time = 300 s), the procedure discontinued (Spreen & Strauss, 1991)

f. Anomalous Sentence Repetition (original Greek version created for the present study based on a description in Lezak et al., 2004), consisting of 14 commonly used sayings which, however, included an incorrect word, to be repeated as given; variable of interest: number of correct items.

g. Graphical Sequence Test (Lezak et al., 2004), consisting of six requests to write or draw specific sequences of objects, shapes, or numbers; variable of interest: number of correct items.

h. Recitation of Months (forwards and backward) (Östberg et al., 2008; Tardif et al., 2013); variables of interest: time to completion forwards, backward, and their ratio (backward: forward).

i. Motor Programming (Lezak et al., 2004), consisting of a congruent (20 trials of tapping the table the same number of times as the examiner, specifically once or twice) and an incongruent (20 trials of tapping the opposite number of times, namely once for two taps and twice for one) condition; variables of interest: number of correct responses on congruent and incongruent conditions.

j. Judgment of Line Orientation (JLO; Benton et al., 1983) (abbreviated for the present study, 10 items, every third item of the original test); variable of interest: number of correct pairs.

k. Clock Drawing Test (CDT), free drawn version (Bozikas et al., 2008; Freedman et al., 1994); variable of interest: total score.

l. Vocabulary test: a Greek multiple-choice vocabulary test (Giaglis et al., 2010); variable of interest: number of correct responses.

Data analyses

Student's *t*-tests were used to compare the mean scores of a group of participants without dementia at baseline assessment, as defined above (CDR = 0 *and* MMSE ≥ 26), who also continued to meet these criteria at follow-up, henceforth referred to as Group 1 (this group was also used in the present analyses to generate normative data) to a group of participants who were without dementia only at baseline examination (CDR = 0 *and* MMSE ≥ 26) but did not meet both of the criteria (i.e., had a CDR > 0 *and* MMSE < 26) at the follow-up assessment, henceforth designated as Group 2. To indicate the size of the differences between the two groups, we also estimated Cohen's effect sizes. For Cohen's *d*, we considered an effect size of 0.2–0.3 as small, at approximately 0.5 medium, and of 0.8–1.0 as a large effect size.

Multiple regression analyses were performed with the neuropsychological test scores as the dependent variables. The independent variables were age at baseline assessment, age squared, education, and sex. The significance of the term of age squared was tested in the model to evaluate nonlinear effects on performance (Parmenter et al., 2010). If an independent variable had an insignificant regression coefficient, it was dropped from the model, and the model was re-estimated, including the remaining variables. The procedure was repeated until all remaining predictors had regression weights significantly different from zero (Oosterhuis et al., 2016). Coefficients of determination (R^2) were reported as a measure of variation that is explained by each model. Statistical analyses were performed using SPSS software version 29.0. [We applied linear, instead of nonlinear, analyses (as proposed by Kornak et al., 2019), to retain the original data, and avoid potential manipulation of data and variables where outliers may exist, as well as because existing software is not readily available for the simultaneous estimation of nonlinear trends]. Traditional norms were derived by dividing our sample into discrete age and educational-level categories and computing the mean (M) and standard deviation (SD), as well as percentile rankings. Performance was classified as normal by utilizing a cutoff score of 1.5 SD, which corresponds to the 7.5th percentile (Petersen & Morris, 2005). Also, raw scores of all variables of interest were

adjusted and converted to scaled standardized scores ($M = 10$, $SD = 3$) using the cumulative frequency distribution.

Finally, we applied each type of normative data to two case examples to illustrate their relative utility in detecting potential cognitive impairment in one patient with known dementia and another with known MCI.

Results

Group comparisons

At baseline, the overall study sample comprised 1984 participants, of which 1089 were also examined in the follow-up. Of the total number of participants who were examined at both assessments, 492 were considered without dementia at both assessments (Group 1) (age range: 65–87 years, education range: 1–21 years) and they constituted the main sample of this study. Of the total sample with baseline and follow-up assessments, 395 participants did not meet criteria for inclusion (based on CDR and MMSE scores) at baseline and were thus excluded from the sample. Finally, 202 participants who were initially classified as without dementia at the baseline assessment, did not meet the criteria at the follow-up assessment (had either CDR > 0 or MMSE < 26); they were included in a comparison group (Group 2) (age range: 65–88 years, education range: 1–18 years) only for the purpose of establishing the need (or lack thereof) for basing normative data on Group 1. Table 1 lists the demographic and clinical characteristics of Groups 1 and 2.

A group comparison of all neuropsychological test variables yielded significantly lower mean scores for Group 2 relative to Group 1 on almost all scores (p-values ranging from < .001–.021; see Table 2) and with moderate effect sizes. Only one very simple neuropsychological test variable [namely, MCGCF-recognition (Cohen's $d = 0.16$, $p > .05$)] did not differentiate the groups from each other (the small effect size suggested no meaningful difference). Overall, these findings indicated that at baseline, Group 2 demonstrated poorer performance than Group 1, suggesting an already existing subtle decline across all tests in the former group several years before the follow-up assessment.

Association between demographic factors and neuropsychological test performance

Henceforth analyses were based only on Group 1. The effects of age, education, and sex were investigated for each test, and significant linear associations were found, as presented in Table 3. As indicated by the adjusted R^2, results showed that about 17% of the variance of test scores included in the neuropsychological battery was explained by these demographic variables, with a range between 4% and 56%. Education was significantly associated with all test scores, with higher education being related to better performance. In contrast, age was not significantly related to CDT, TMT B, ratio of TMT B to TMT A, vocabulary, comprehension, recitation of months, phonological verbal fluency, visual recognition memory, motor programing–congruent condition, the 1st learning trial of GVLT and two of the deficit indices of the GVLT (consolidation and retrieval deficit). For the remaining test variables, greater age was related to poorer performance. Sex was significantly related to test scores on the GVLT, MCG (for copy and delayed recall scores but not for immediate recall), JLO, CDT, verbal fluency, and Anomalous Sentence Repetition, but not with any other test variables. Specifically, we found a female advantage on GVLT learning and recall scores, the copy condition of the MCG, verbal fluency (both phonological and semantic), and

Table 1. Demographic and clinical characteristics during the baseline evaluation for Group 1 and Group 2

Characteristic	Group 1 $n = 492$	Group 2 $n = 202$	p-value
Demographic			
Age, yrs (SD)	71.5 (4.3)	73.6 (5.2)	< .001
Education, yrs (SD)	10.4 (4.7)	7.4 (3.9)	< .001
Male sex, n (%)	204 (41.5)	80 (39.6)	.651
Clinical			
GDS total score, mean (SD)	1.66 (2.8)	1.9 (3.3)	.166
HADS total score, mean (SD)	2.3 (3.4)	2.6 (4.0)	.152
Hypertension, n (%)	310 (63.0)	120 (59.4)	.502
Diabetes, n (%)	68 (13.8)	29 (14.4)	.839
Coronary Disease, n (%)	43 (8.7)	16 (7.9)	.735
Presence of ε4 allele, n (%)	50 (10.2)	24 (11.9)	.894

GDS = Geriatric Depression Scale, HADS = Hospital Anxiety and Depression Scale. Chi-squared test used for categorical variables; t-tests for continuous variables. Data on continuous variables are presented as $M \pm SD$; data on categorical variables are presented as number of occurrences.

Anomalous Sentence Repetition, but a male advantage on MCG immediate recall percentage, delayed recall, delayed recall percentage, and JLO. No other significant sex effects were found. The term of age squared was not significant in any of the regression equations, thus it was not included in further analyses.

Calculation of regression-based normative equations

Normative data are presented as regression-based algorithms to adjust test scores for age, education, and sex variables according to the following equation:

$$\text{Test Score}_{predicted} = \text{constant} + \beta_{age}(\text{age}) + \beta_{sex}(\text{sex}) + \beta_{education}(\text{education})$$

To determine size of deviation from the predicted score, we calculated the following:

$$(\text{Test score}_{actual} - \text{Test score}_{predicted})/SD \text{ residual}$$

Individual equations are listed in Table 3 but are also available in an Excel file for clinical use for individual examinees [available in compressed form (https://www.7-zip.org/) at https://www.psy.auth.gr/faculty/kosmidis/NormsCalculationFINAL.7z). Based on this procedure, we present two case examples. In the case of A, a 76-year-old woman with 6 years of education and a diagnosis of dementia, performance on the GVLT sum of five learning trials can be calculated given the information available in Table 3. Specifically, the expected score = $44.135 + (-0.212 \times 76) + (7.478 \times 1) + (0.752 \times 6) = 40.013$. Instead, her actual score was 15, clearly below what was predicted based on her demographic characteristics. To determine whether or not the actual score indicated impairment, we continued as follows: $(15 - 40.013) / 10.04 = -2.49$, suggesting that, indeed, the difference was significant. Similarly, for Case B, a 74-year-old man with 2 years of education and a diagnosis of MCI, had a predicted score = $44.135 + (-0.212 \times 74) + (7.478 \times 0) + (0.752 \times 2) = 29.951$; yet his actual score was also 15, clearly below that predicted and confirmed a significantly lower as follows: $(15 - 29.951)/10.04 = -1.49$.

Calculation of discrete normative data: raw and standard scores

Discrete normative data, with corresponding raw test scores, mean (SD), and percentiles, are presented in table format in Appendix A.

Table 2. Group comparisons (means and SDs) on neuropsychological test variables and between-group effect sizes (Cohen's *d*) (continued)

Test scores	Group 1 $n = 492$		Group 2 $n = 202$				Effect size
	Mean	SD	Mean	SD	*t*	*p*-value	Cohen's *d*
Memory							
GVLT							
1st learning trial	6.13	2.06	4.99	2.04	6.653	<.001	0.56
5th learning trial	12.00	2.77	10.28	2.74	7.449	<.001	0.62
Total Sum of learning trials	48.63	11.30	40.85	10.48	8.370	<.001	0.71
Learning curve	1.40	0.58	1.25	0.60	3.085	.001	0.25
Immediate free recall	10.11	3.27	7.74	3.09	8.782	<.001	0.74
Immediate cued recall	12.10	2.59	10.34	2.52	8.139	<.001	0.69
Delayed free recall	10.80	3.33	8.45	3.41	8.321	<.001	0.70
Delayed cued recall	12.19	2.69	10.38	2.74	7.964	<.001	0.67
Recognition total score	14.73	1.68	14.29	1.96	2.784	.003	0.24
ISDA encoding deficit	5.37	3.12	7.44	2.98	−7.295	<.001	0.68
ISDA consolidation deficit	0.12	0.13	0.18	0.13	−5.363	<.001	0.46
ISDA retrieval deficit	0.38	0.19	0.49	0.18	−6.179	<.001	0.59
MCGCF							
Copy	34.51	2.80	33.62	3.45	3.513	<.001	0.29
Copy percentage (% 36)	95.86	7.78	93.39	9.58	3.514	<.001	0.28
Immediate recall	19.48	7.21	15.40	6.82	6.788	<.001	0.58
Immediate recall percentage (% copy)	56.03	19.50	45.55	18.99	6.402	<.001	0.54
Delayed recall	18.03	7.34	13.69	7.02	7.042	<.001	0.60
Delayed recall percentage (% copy)	51.83	20.03	40.43	19.79	6.719	<.001	0.57
Recognition	17.97	2.30	17.62	2.16	.744	.229	0.16
Visuospatial Perception							
Judgment of Line Orientation	6.32	2.19	5.37	2.40	5.831	<.001	0.41
Clock Drawing Test	13.09	2.09	12.25	2.69	2.987	.002	0.35
Attention/Processing Speed							
Trail Making Test							
Part A	65.48	29.85	84.55	39.52	−6.063	<.001	0.54
Part B	160.47	86.04	221.32	112.01	−5.932	<.001	0.61
Part B:Part A	2.60	1.05	3.06	1.38	−3.565	<.001	0.38
Language							
Boston Naming-Short form	12.21	2.48	10.57	2.50	7.862	<.001	0.62
Comprehension	11.44	1.08	10.83	1.50	6.006	<.001	0.47
Repetition	5.25	1.07	4.85	1.15	4.422	<.001	0.36
Verbal fluency							
Semantic (objects)	17.82	4.61	15.00	5.03	6.828	<.001	0.58
Phonological (letter "A")	9.04	4.82	6.82	4.15	6.079	<.001	0.49
Executive functioning							
Months test							
Forward	6.74	2.08	8.04	5.16	−2.062	.021	0.33
Backward	14.85	8.31	24.04	22.26	−3.353	<.001	0.55
Backward:Forward ratio	2.33	1.46	3.36	3.10	−2.659	.005	0.43
Repetition of anomalous sentences	13.16	1.39	12.47	1.60	4.484	<.001	0.46
Graphical sequence	5.14	0.90	4.70	1.05	4.382	<.001	0.46
Motor programing							
Congruent	19.65	0.89	19.40	1.24	2.132	.017	0.23
Incongruent	19.29	1.90	18.78	2.55	2.467	.018	0.23
Intellectual functioning							
Vocabulary	26.83	11.31	18.79	10.28	8.620	<.001	0.74

T-test used for all continuous variables.

These data are stratified by four levels of age (65–69, 70–74, 75–79, and 80–87) and by three levels of education (1–6, 7–12, 13+ years). Where sex had an effect on performance, separate tables for men and women were included. All tables include means and standard deviations by cell, percentile rankings, and cutoff scores for easy comparison to a target patient. These tables include norms based on raw data. While for most test variables, individual scores corresponded to a specific percentile ranking, on a few variables, the same score corresponded to several percentile rankings. For example, for both women and men in some age/education cells, a score of 100% (or 36) on the MCG copy condition, as well as on the Sentence Comprehension of the BDAE, corresponded with performance at the 25th, 50th, the 75th, and the 90th percentile (see Appendix A).

Using our previous case examples, Case A's score of 15 on the GVLT sum of five learning trials for women would be −2.16 SD below the mean for her age and educational-level group (on page 7 of Appendix A, see the eighth column from the left on the relevant table), which is also below the cutoff for determining impaired performance (below the 7.5th percentile). For Case B, his score of 15 on the same variable was also below the 7.5th percentile (see the fifth column from the left on the relevant table) and −2.92 SD below the mean for men in his age group and educational level, thus also impaired.

Standard scores (SS) were calculated by taking the raw score of each variable of interest and transforming it to a common scale ($M = 10$, $SD = 3$) using the cumulative frequency distribution. This approach allows for the organization and presentation of normative

Table 3. Regression-based equations to calculate adjusted scores

Test scores	F (df)	R^2	SE_{est}^a	Normative equation[b]	p-value[c]
Memory					
GVLT					
1st learning trial	13.95 (2, 486)	.05	2.01	4.214 + (0 * age) + (.711 * sex) + (.076 * education)	< .001
5st learning trial	52.48 (3, 484)	.25	2.42	11.744 + (−.0667 * age) + (1.809 * sex) + (.221 * education)	< .001
Total Sum of learning trials	43.15 (3, 484)	.21	10.01	44.135 + (−.212 * age) + (7.478 * sex) + (.752 * education)	< .001
Learning curve	26.01 (3, 484)	.14	0.55	1.821 + (−.016 * age) + (.267 * sex) + (.032 * education)	< .001
Immediate free recall	49.64 (3, 484)	.24	2.87	8.687 + (−.064 * age) + (2.396 * sex) + (.216 * education)	< .001
Immediate cued recall	30.36 (3, 484)	.16	2.38	12.633 + (−.101 *age) + (1.625 * sex) + (.145 * education)	< .001
Delayed free recall	49.86 (3, 485)	.24	2.92	9.421 + (−.066 * age) + (2.328 * sex) + (.235 * education)	< .001
Delayed cued recall	53.59 (3, 485)	.18	2.44	7.557 + (0 * age) + (1.874 *sex) + (.160 * education)	< .001
Recognition total score	16.56 (3, 485)	.06	1.62	13.034 + (0 * age) + (.639 * sex) + (.066 * education)	< .001
ISDA encoding deficit	34.81 (2, 449)	.19	2.82	5.204 + (.072 * age) + (−1.960 * sex) + (−.180 * education)	.035
ISDA consolidation deficit	8.61 (2, 450)	.04	0.13	.218+ (0 * age) + (−.046 * sex) + (−.003 * education)	< .001
ISDA retrieval deficit	27.41 (2, 450)	.11	0.18	.639 + (0 * age) + (−.098 * sex) + (−.010 * education)	< .001
MCGCF					
Copy	22.57 (3, 483)	.12	2.62	39.520 + (−.107 * age) + (.583 * sex) + (.164 * education)	< .001
Copy percentage (% 36)	22.57 (3, 483)	.12	7.30	109.782 + (−.296 * age) + (1.621 * sex) + (.455 * education)	< .001
Immediate recall	50.64 (2, 483)	.17	6.57	28.918 + (−.218 * age) + (0 * sex) + (.597 * education)	< .001
Immediate recall percentage (% copy)	29.17 (3, 482)	.15	17.99	89.129 + (−.589 * age) + (−3.796 * sex) (1.458 * education)	< .001
Delayed recall	38.71 (3, 481)	.19	6.61	35.569 + (−.300 * age) + (−1.469 * sex) + (.604 * education)	< .001
Delayed recall percentage (% copy)	33.42 (3, 481)	.17	18.28	97.039 + (−.746 * age) + (−4.852 *sex) + (1.534 * education)	< .001
Recognition	7.04 (2, 151)	.09	2.21	18.476 + (0 * age) + (−.989 * sex) + (.096 * education)	.001
Visuospatial Perception					
Judgment of Line Orientation	19.39 (3, 481)	.11	2.02	11.147 + (−.058 * age) + (−.970 * sex) + (.098 * education)	< .001
Clock Drawing Test	9.09 (2, 322)	.05	1.99	11.284 + (0 * age) + (.567 * sex) + (.087 * education)	< .001
Attention/Processing Speed					
Trail Making Test					
Part A	78.39 (2, 479)	.25	24.01	−[35.049 + (1.738 * age) + (0 * sex) + (−2.374 * education)]	< .001
Part B	153.558 (2, 405)	.28	52.62	224.075 + (0 * age) + (0 * sex) + (−7.190 * education)]	< .001
Part B:Part A	32.31 (1, 403)	.07	0.91	3.140 + (0 * age) + (0 *sex) + (−.057 * education)	.001
Language					
Boston Naming−Short form	161.98 (2, 486)	.40	1.93	16.077 + (−.100 * age) + (0 * sex) + (.317 * education)	< .001
Comprehension	63.82 (2, 484)	.12	1.02	10.619 + (0 * age) + (0 * sex) + (.080 * education)	< .001
Repetition	66.45 (2, 486)	.22	0.95	6.319 + (−.030 * age) + (0 * sex) + (.101 * education)	< .001
Verbal fluency					
Semantic	16.01 (3, 485)	.09	4.41	22.569 + (−.127 * age) + (1.293 *sex) + (.220 * education)	< .001
Phonological	101.63 (2, 485)	.30	4.01	2.035 + (0 * age) + (.740 * sex) + (.563 * education)	.008
Executive functioning					
Months test					
Forward	44.75 (1, 311)	.32	1.95	8.673 + (0 * age) + (0 * sex) + (−.165 * education)	< .001
Backward	44.53 (1, 311)	.13	7.78	22.583 + (0 * age) + (0 * sex) + (−.657 * education)	< .001
Backward:Forward ratio	7.57 (1, 310)	.02	1.44	2.926 + (0 * age) + (0 * sex) + (−.050 * education)	< .001
Repetition of anomalous sentences	55.09 (3, 429)	.20	1.25	15.789 + (−.062 * age) + (.472 *sex) + (.100 * education)	< .001
Graphical sequence	37.45(2, 429)	.15	0.83	6.970 + (−.035 * age) + (0 * sex) + (.064 * education)	< .001
Go/no−Go					< .001
Congruent	24.68 (1, 432)	.05	0.86	19.160 + (0 * age) + (0 * sex) + (.045 * education)	< .001
Incongruent	16.78 (1, 427)	.04	1.86	18.418 + (0 * age) + (0 * sex) + (.081 * education)	< .001
Intellectual functioning					< .001
Vocabulary	578.67 (1, 462)	.56	7.41	7.676 + (0 * age) + (0 * sex) + (1.819 * education)	< .001

All F tests are significant at $p < .001$. Subtest scores are raw scores.
[a]Standard error of the estimate.
[b]Age is in years; gender is coded as 0=male. 1=female; Male sex is used as reference; education is coded in educational years.
[c]$\beta = 0$ if $p > 0.05$.

data and facilitates comparisons between different neuropsychological tests. Conversion of raw scores to standard scores is listed in table format in Appendix B. Again, referring to the previous case examples, for both Case A and Case B, their GVLT sum of learning trials standard score would be below $SS = 2$ (in the low range).

Discussion

In the present study, we found that ensuring that normative samples of older individuals are demonstrably untainted by an early neurodegenerative process yielded stricter criteria, as mean values were higher in those demonstrably without dementia at follow-up than those who upon follow-up no longer were deemed to be in this category. This was the case for almost all neuropsychological test variables, with the exception of one very simple task. More specifically, we found that, at baseline assessment, Group 1 performed better than Group 2 on almost all tests. This finding justified our efforts to secure such an untainted group as a robust normative standard, as it suggests that individuals in Group 2 were already presenting subtle cognitive decline, a few years before their diagnosis of MCI or dementia. We also found an education effect on all test variables, an age effect on some, and a sex effect on a few. Thus, we included these demographic characteristics in the calculation of normative data according to commonly used approaches. Regression-based normative data appeared to be more sensitive to the level of impairment based on demographic factors, than discrete norms. Specifically, the former differentiated the level of performance of two cases (one with diagnosed MCI and at the impairment cutoff, and the other with diagnosed dementia and clearly deficient) with

the same score, but different ages, sexes, and years of education. In contrast, when comparing these same cases to discrete norms tables, they were both below the cutoff percentile and in the deficient range based on cell means and SDs. Moreover, standardized scaled scores overestimated the severity of impaired performance, most likely because they were based on the total Group 1 sample and did not consider the effects of age, sex, and education. This highlights the importance of accounting for factors that influence cognitive performance when developing norms, as well as the preferability of regression-based norms.

Older adults present greater variability in their neuropsychological performance than younger cohorts (Duff et al., 2010) and more rapid rates of cognitive decline with aging (Hänninen et al., 1996), thus necessitating normative data adjusted for this population to distinguish normal changes from those reflective of an underlying neurodegenerative disease. Through the procedure described in the present study, we addressed frequent criticisms related to the interpretation of neuropsychological performance of healthy older individuals, questioning the validity of data that may be, unbeknownst to both experimenter and research participant, tainted by undiagnosed early neurocognitive decline. By studying individuals from a population-based epidemiological study, we ensured our sample's representativeness of the population. By choosing among these individuals, those who were considered to be without MCI or dementia on two assessment times several years apart, we confirmed that our sample was not in the early, and yet undetected, phase of a neurodegenerative process. Thus, our sample provided a unique opportunity to ensure the *ex post facto* exclusion of preclinical MCI or dementia cases from the normative data within a population-based large sample of older individuals. Excluding individuals from the final normative sample whose cognitive status no longer justified their characterization as without MCI or dementia at follow-up yielded stricter criteria for the determination of cognitive impairment than typically done in most normative studies.

Additional concerns relative to the accuracy and utility of normative data, especially for older individuals, relate to the method of norm development. Given the more accelerated rate of cognitive decline typically observed among healthy older individuals relative to their younger counterparts, we maximized the sensitivity of our normative data to the effects of demographic characteristics often found to affect neuropsychological performance. Thus, we explored the influence of age, education, and sex on neuropsychological performance on each test variable. Indeed, we found that these characteristics were associated with many neuropsychological variables. Specifically, education was associated with performance across the board, wherein increased education was related to better neuropsychological performance. Age was associated with most variables (verbal and visual memory and learning, visuospatial perception, language, attention/processing speed, and executive functioning), but not all; where an association existed, it was inverse, with increasing age being related to poorer performance. Finally, the association of sex with neuropsychological performance was typical of that in the literature (female advantage on verbal learning and recall, figure copy, attention/processing speed, verbal fluency, and inhibition, and male advantage on visual memory (immediate recall percentage, delayed recall, and delayed recall percentage) and visuospatial perception of orientation). Consequently, these demographic variables were included in the analyses when developing the regression-based normative data and were also used in the discrete norms to break down our sample into five-year age groups, three educational-level groups and sex (where indicated by group differences).

In the course of exploring the optimal procedure for enhancing the utility of normative data, we co-normed a comprehensive battery of commonly used neuropsychological tests based on this sample of older adults without dementia. Such studies are limited in number, especially for this age group (Duff, 2016). For example, a study in New Zealand (Callow et al., 2015) developed normative data for a large sample of individuals aged 45–85, who were tested at two-time points. This, however, was limited to Addenbrooke's Cognitive Examination-Revised, a brief screening test. Another study (Roberts et al., 2008), conducted in the US, developed normative data for ten commonly used neuropsychological tests, but only for two relatively broad age groups: 70–79 and 80–89 years old. Both studies included some participants who were tested over the telephone, and produced traditional, discrete norms. In contrast, our co-norming of a comprehensive battery of tests may aid in process-based interpretation of neuropsychological performance, as clinicians and researchers can thus view and evaluate an individual's scores on one test or cognitive domain relative to another, and, accordingly, make sense of the individual's overall pattern of performance and its potential consistency with (or lack thereof) expected patterns of cognitive strengths and weaknesses. Thus, the present neuropsychological battery may be used flexibly in clinical practice.

Our study has some limitations. Total variance in our regression models varied and accounted for a mean value of about 17%. The magnitude of this correction indicated that other demographic or non-demographic variables might also be potential predictors of test performance. Another issue to be addressed concerning both types of norms relates to the small number of individuals in the normative group who were 80 years of age or older, which resulted from our sample's age distribution. For the regression-based norms, this would reduce precision; for the discrete norms, it may lead to overestimating or underestimating cognitive decline among the very old. Moreover, despite our large sample, on several tests discrete norms yielded scores that were equated to more than one percentile level. This raises concerns regarding the interpretation of performance levels, as well as to the sensitivity of the particular variables in differentiating abilities. Finally, we compared types of norms using only two individual case examples and one specific neuropsychological test variable, thus limiting our ability to generalize to other test variables. A larger sample and an extensive list of test variables should be used to confirm the preferability of one type of norms over the other for each test variable, as these may vary either in a systematic or even a nonsystematic way.

The current study presents a method for ensuring the relative purity of normative data for a comprehensive, flexible neuropsychological test battery comprised of well-known tests applied in everyday clinical practice and research studies, specifically for older individuals. This has the potential to increase the accuracy in identifying individuals who are even in the early, and yet undiagnosed, stage of a neurodegenerative process. While our data may be specific to Greek-speaking populations, they provide a template for similar future studies in other languages and cultures. Additionally, they may aid in cross-national studies for comparisons in international collaborations. In any event, the investigation and refinement of approaches that augment the accuracy of neuropsychological assessment for the aging population, such as differentiating the effects of normal aging and early neurodegenerative processes, is certainly a necessary and worthwhile endeavor.

References

Alden, E. C., Lundt, E. S., Twohy, E. L., Christianson, T. J., Kremers, W. K., Machulda, M. M., Jack, C. R., Jr, Knopman, D. S., Mielke, M. M., Petersen, R. C., & Stricker, N. H. (2022). Mayo normative studies: A conditional normative model for longitudinal change on the auditory verbal learning test and preliminary validation in preclinical alzheimer's disease. *Alzheimer's & Dementia (Amsterdam, Netherlands)*, *14*(1), e12325. https://doi.org/10.1002/dad2.12325

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. American Psychiatric Association.

Benton, A. L., Hamsher, K. D., Varney, N., & Spreen, O. (1983). *Contributions to Neuropsychological Assessment: A Clinical Manual*. Oxford University Press.

Bougea, A., Maraki, M. I., Yannakoulia, M., Stamelou, M., Xiromerisiou, G., Kosmidis, M. H., Ntanasi, E., Dardiotis, E., Hadjigeorgiou, G., Sakka, M., Anastasiou, P., C., A., Stefanis, L., & Scarmeas, N. (2019). Higher probability of prodromal parkinson's disease is related to lower cognitive performance. *Neurology*, *92*(19), e2261–e2272. https://doi.org/10.1212/WNL.0000000000007453

Bozikas, V. P., Giazkoulidou, A., Hatzigeorgiadou, M., Karavatos, A., & Kosmidis, M. H. (2008). Do age and education contribute to performance on the clock drawing test? Normative data for the greek population. *Journal of Clinical and Experimental Neuropsychology*, *30*(2), 199–203. https://doi.org/10.1080/13803390701346113

Brandt, J., & Benedict, R. H. B. (2001). *Hopkins verbal learning test-revised. Professional manual*. Psychological Assessment Resources, Inc.

Callow, L., Alpass, F., Leathem, J., & Stephens, C. (2015). Normative data for older New Zealanders on the Addenbrooke's cognitive examination-revised. *New Zealand Journal of Psychology*, *44*(3), 29–41.

Dardiotis, E., Kosmidis, M. H., Yannakoulia, M., Hadjigeorgiou, G. M., & Scarmeas, N. (2014). The hellenic longitudinal investigation of aging and diet (HELIAD): Rationale, study design, and cohort description. *Neuroepidemiology*, *43*(1), 9–14. https://doi.org/10.1159/000362723

Duff, K. (2016). Demographically corrected normative data for the Hopkins verbal learning test-revised and brief visuospatial memory test-revised in an elderly sample. *Applied Neuropsychology*, *23*(3), 179–185. https://doi.org/10.2144/000114329

Duff, K., Beglinger, L. J., Moser, D. J., Paulsen, J. S., Schultz, S. K., & Arndt, S. (2010). Predicting cognitive change in older adults: The relative contribution of practice effects. *Archives of Clinical Neuropsychology*, *25*(2), 81–88. https://doi.org/10.1093/arclin/acp105

Duff, K., & Ramezani, A. (2015). Regression-based normative formulae for the repeatable battery for the assessment of neuropsychological status for older adults. *Archives of Clinical Neuropsychology*, *30*(7), 600–604. https://doi.org/10.1093/arclin/acv052

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

Freedman, M., Leach, L., Kaplan, E., Winocur, G., Shulman, K. I., & Delis, D. C. (1994). *Clock drawing: A neuropsychological analysis. Clock drawing: A neuropsychological analysis*. Oxford University Press.

Giaglis, G., Kyriazidou, S., Paraskevopoulou, E., Tascos, N., & Kosmidis, M. H. (2010). Evaluating premorbid level: Preliminary findings regarding the vulnerability of scores on cognitive measures in patients with MS. *Journal of the International Neuropsychological Society*, *16*(Suppl 1), 159.

Hänninen, T. U. O. M. O., Koivisto, K. E. I. J. O., Reinikainen, K. A. R. I. J., Helkala, E. E. V. A..-L. I. I. S. A., Soininen, H. I. L. K. K. A., Mykkänen, L. E. E. N. A., Laakso, M. A. R. K. K. U., & Riekkinen, P. A. A. V. O. J. (1996). Prevalence of ageing-associated cognitive decline in an elderly population. *Age and Ageing*, *25*(3), 201–205. https://doi.org/10.1093/ageing/25.3.201

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults.* Psychological Assessment Resources, Inc.

Kornak, J., Fields, J., Kremers, W., Farmer, S., Heuer, H. W., Forsberg, L., Brushaber, D., Rindels, A., Dodge, H., Weintraub, S., Besser, L., Appleby, B., Bordelon, Y., Bove, J., Brannelly, P., Caso, C., Coppola, G., Dever, R., Dheel, C., Dickerson, B., Dickinson, S., Dominguez, S., Domoto-Reilly, K., Faber, K., Ferrall, J., Fishman, A., Fong, J., Foroud, T., Gavrilova, R., Gearhart, D., Ghazanfari, B., Ghoshal, N., Goldman, J., Graff-Radford, J., Graff-Radford, N., Grant, I. M., Grossman, M., Haley, D., Hsiao, J., Hsiung, R., Huey, E. D., Irwin, D., Jones, D., Jones, L., Kantarci, K., Karydas, A., Kaufer, D., Kerwin, D., Knopman, D., Kraft, R., Kramer, J., Kukull, W., Lapid, M., Litvan, I., Ljubenkov, P., Lucente, D., Lungu, C., Mackenzie, I., Maldonado, M., Manoochehri, M., McGinnis, S., McKinley, E., Mendez, M., Miller, B., Multani, N., Onyike, C., Padmanabhan, J., Pantelyat, A., Pearlman, R., Petrucelli, L., Potter, M., Rademakers, R., Ramos, E. M., Rankin, K., Rascovsky, K., Roberson, E. D., Rogalski-Miller, E., Sengdy, P., Shaw, L., Staffaroni, A. M., Sutherland, M., Syrjanen, J., Tartaglia, C., Tatton, N., Taylor, J., Toga, A., Trojanowski, J., Wang, P., Wong, B., Wszolek, Z., Boeve, B., Boxer, A., Rosen, H., & ARTFL/LEFFTDS Consortium (2019). Nonlinear Z-score modeling for improved detection of cognitive abnormality. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*(1), 797–808. https://doi.org/10.1016/j.dadm.2019.08.003

Kosmidis, M. H., Vlachos, G., Dardiotis, E., Yannakoulia, M., Hadjigeorgiou, G. M., & Scarmeas, N. (2018). Prevalence of dementia and its subtypes in Greece. *Alzheimer Disease & Associated Disorders-An International Journal*, *32*(3), 232–239. https://doi.org/10.1097/WAD.0000000000000249

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., & Kiosseoglou, G. (2004). The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society*, *10*(2), 164–172. https://doi.org/10.1017/S1355617704102014

Kramer, A. O., Casaletto, K. B., Umlauf, A., Staffaroni, A. M., Fox, E., You, M., & Kramer, J. H. (2020). Robust normative standards for the California verbal learning test (CVLT) ages 60-89: A tool for early detection of memory impairment. *The Clinical Neuropsychologist*, *34*(2), 384–405. https://doi.org/10.1080/13854046.2019.1619838

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th edn). Oxford University Press.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th edn). Oxford University Press.

Mandyla, M. A., Yannakoulia, M., Hadjigeorgiou, G., Dardiotis, E., Scarmeas, N., & Kosmidis, M. H. (2022). Identifying appropriate neuropsychological tests for uneducated/Illiterate older individuals. *Journal of the International Neuropsychological Society*, *28*(8), 862–875. https://doi.org/10.1017/S1355617721001016

Martin, R., Sawrie, S., Gilliam, F., Mackey, M., Faught, E., Knowlton, R., & Kuzniekcy, R. (2002). Determining reliable cognitive change after epilepsy surgery: Development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. *Epilepsia*, *43*(12), 1551–1558. https://doi.org/10.1046/j.1528-1157.2002.23602

McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*(3), 203–221. https://doi.org/10.1007/BF02214762

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, *34*(7), 939–944.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd edn). Oxford University Press.

Morris, J. C. (1993). The clinical dementia rating (CDR). *Neurology*, *43*(11), 2412–2412. https://doi.org/10.1212/WNL.43.11.2412-a

Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191–202. https://doi.org/10.1177/1073191115580638

Östberg, P., Fernaeus, S. E., Bogdanović, N., & Wahlund, L. O. (2008). Word sequence production in cognitive decline: Forward ever, backward never. *Logopedics Phoniatrics Vocology*, *33*(3), 126–135. https://doi.org/10.1080/14015430801945794

Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. B. (2010). The utility of regression-based norms in interpreting the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, *16*(1), 6–16. https://doi.org/10.1017/S1355617709990750

Petersen, R. C., & Morris, J. C. (2005). Mild cognitive impairment as a clinical entity and treatment target. *Archives of Neurology*, *62*(7), 1160–1163. https://doi.org/10.1001/archneur.62.7.1160

Roberts, R. O., Geda, Y. E., Knopman, D. S., Cha, R. H., Pankratz, V. S., Boeve, B. F., Ivnik, R. J., Tangalos, E. G., Petersen, R. C., & Rocca, W. A. (2008). The Mayo clinic study of aging: Design and sampling, participation, baseline measures, and sample characteristics. *Neuroepidemiology*, *30*(1), 58–69. https://doi.org/10.1159/000115751

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change timothy. *Neuropsychology*, *22*(6), 800–811. https://doi.org/10.1037/a0013091

Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd edn). Oxford University Press.

Tardif, M., Roy, M., Remi, B., Laforce, R., Verret, L., Fortin, M-Pierre, Houde, M., & Poulin, S. (2013). P4-103: Months backward test as a reliable predictor of cognitive decline in mild Alzheimer's disease. *Alzheimer's & Dementia*, *9*(4), P741–P742. https://doi.org/10.1016/j.jalz.2013.05.1493

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*(4), 357–369. https://doi.org/10.1017/S1355617799544068

Tsapkini, K., Vlahou, C. H., & Potagas, C. (2010). Adaptation and validation of standardized aphasia tests in different languages: Lessons from the Boston diagnostic aphasia examination-short form in greek. *Behavioural Neurology*, *22*(3-4), 111–119. https://doi.org/10.3233/BEN-2009-0256

Vlachos, G. S., Kosmidis, M. H., Yannakoulia, M., Dardiotis, E., Hadjigeorgiou, G., Sakka, P., Ntanasi, E., Stefanis, L., & Scarmeas, N. (2020). Prevalence of mild cognitive impairment in the elderly population in Greece: Results from the HELIAD study. *Alzheimer Disease & Associated Disorders*, *34*(2), 156–162. https://doi.org/10.1097/WAD.0000000000000361

Vlahou, C. H., & Kosmidis, M. H. (2002). The greek trail making test: Preliminary norms for clinical and research use. *Psychology: The Journal of the Hellenic Psychological Society (in Greek)*, *9*(3), 336–352.

Vlahou, C. H., Kosmidis, M. H., Dardagani, A., Tsotsi, S., Giannakou, M., Giazkoulidou, A., Zervoudakis, E., & Pontikakis, N. (2013). Development of the greek verbal learning test: Reliability, construct validity, and normative standards. *Archives of Clinical Neuropsychology*, *28*(1), 52–64. https://doi.org/10.1093/arclin/acs099