Early detection of Alzheimer's disease with a total score of the German CERAD

MICHAEL M. EHRENSPERGER,[1] MANFRED BERRES,[2] KIRSTEN I. TAYLOR,[1,3] AND
ANDREAS U. MONSCH[1]

[1]Department of Geriatrics, University Hospital Basel, Basel, Switzerland
[2]Department of Mathematics and Technology, University of Applied Sciences, Koblenz, Germany
[3]Department of Experimental Psychology, University of Cambridge, Cambridge, United Kingdom

Abstract

The goal of the present study was to evaluate the diagnostic discriminability of three different global scores for the German version of the Consortium to Establish a Registry on Alzheimer's Disease-Neuropsychological Assessment Battery (CERAD-NAB). The CERAD-NAB was administered to 1100 healthy control participants [NC; Mini-Mental State Examination (MMSE) mean = 28.9] and 352 patients with very mild Alzheimer's disease (AD; MMSE mean = 26.1) at baseline and subsets of participants at follow-up an average of 2.4 (NC) and 1.2 (AD) years later. We calculated the following global scores: Chandler et al.'s (2005) score (summed raw scores), logistic regression on principal components analysis scores (PCA-LR), and logistic regression on demographically corrected CERAD-NAB variables (LR). Correct classification rates (CCR) were compared with areas under the receiver operating characteristics curves (AUC). The CCR of the LR score (AUC = .976) exceeded that of the PCA-LR, while the PCA-LR (AUC = .968) and Chandler (AUC = .968) scores performed comparably. Retest data improved the CCR of the PCA-LR and Chandler (trend) scores. Thus, for the German CERAD-NAB, Chandler et al.'s total score provided an effective global measure of cognitive functioning, whereby the inclusion of retest data tended to improve correct classification of individual cases. (*JINS*, 2010, *16*, 910–920.)

Keywords: Dementia, Cognition, Longitudinal study, Neuropsychology, Neuropsychological test, Change score

INTRODUCTION

Global measures of cognitive functioning are important in both the clinical and research settings. They allow clinicians to efficiently gauge a patient's overall level of functioning, and to communicate this information to other clinicians. In research studies, global measures of cognitive status provide investigators and readers a practical yardstick with which to judge participants' disease severity. Global measures are particularly useful in cross-sectional and longitudinal intervention studies, where complex patterns of performance on different cognitive subtests and across individuals, and variable rates of changes on different cognitive subtests and across individuals, would otherwise be difficult to assess, scrutinize, and communicate. Thus, although global performance measures are necessarily associated with a loss of information which would be unacceptable in the diagnostic and differential diagnostic process, they represent an elegant solution to a variety of practical issues in the clinical and research environment.

Pharmacologic and non-pharmacologic intervention studies have, therefore, used simple and simplifying measures of global functioning to judge cognitive performance at baseline and over time. These include in particular the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), the Alzheimer Disease Assessment Scale (ADAS-Cog; Rosen, Mohs, & Davis, 1984), and for severely impaired individuals the Severe Impairment Battery (SIB; Panisset, Roundier, Saxton, & Boller, 1994). The Consortium to Establish a Registry on Alzheimer's Disease-Neuropsychological Assessment Battery (CERAD-NAB; Morris, Mohs, Rogers, Fillenbaum, & Heyman, 1988; Welsh et al., 1994) may provide the basis for an alternative global assessment measure. The CERAD-NAB has proven useful for the diagnosis of Alzheimer's disease (AD), and a validated and normed version of this test battery has been established

Correspondence and reprint requests to: Michael M. Ehrensperger University Hospital Basel, Memory Clinic, Schanzenstrasse 55, 4031 Basel, Switzerland. E-mail: ehrenspergerm@uhbs.ch

in German-speaking Europe as a minimal common screening battery for dementia syndromes (German version of the CERAD-NAB; available at Memory Clinic, Basel, Switzerland, http://www.memoryclinic.ch; Aebi, 2002; Thalmann et al., 2000). The CERAD has also proven sensitive to cognitive impairments occurring in the early stages of dementia and to cognitive changes over long time spans (Fillenbaum, Unverzagt, Ganguli, Welsh-Bohmer, & Heyman, 2002; Morris et al., 1989, 1993; Welsh, Butters, Hughes, Mohs, & Heyman, 1991, 1992; Zehnder, Bläsi, Berres, Spiegel, & Monsch, 2007).

Chandler et al. (2005) recently developed a total score for the American version of the CERAD-NAB. This total score was created by summing six CERAD subtest scores (excluding MMSE and praxis recall; maximum score = 100) and submitting the sum to a regression analysis to correct for demographic status. They provide normative data (*T*-scores) for these demographically corrected CERAD-NAB total scores. Thus, this score has the advantage that it is easy to calculate and characterizes global cognitive performance within the patient's demographic framework. As expected, this score discriminated normal control participants (NC) from patients with probable AD significantly better than the MMSE (Chandler et al., 2005). However, it is unclear whether the simple sum of CERAD-NAB variables provides the most powerful diagnostic discrimination between NC and AD. Moreover, it remains to be established whether Chandler et al.'s and other potential composite scores of CERAD-NAB performance retain their discriminatory power with AD individuals in very early stages of dementia as well during longitudinal assessment.

The goals of the present study were, therefore, to (1) determine the discriminatory utility of Chandler et al.'s total score of German CERAD-NAB performance with groups of NC participants and AD patients in very mild stages of the disease (MMSE in both groups ≥ 24/30), (2) determine the relative diagnostic discriminabilities of three different global CERAD-NAB scores with groups of NC individuals and AD patients, (3) to determine and compare the longitudinal discriminatory power of the different global CERAD-NAB scores, and (4) examine whether a combination of cross-sectional and longitudinal scores improves diagnostic classification.

MATERIALS AND METHODS

Healthy Aged Participants

A total of 1,100 healthy aged individuals (NC) participated (see Table 1). These individuals were a subset of BASEL study participants (Basel Study on the Elderly; Monsch et al., 2000) and formed the normative sample for the German version of the CERAD-NAB. Baseline testing took place between 1997 and 2001. Additional neuropsychological tests included the following: Trail Making Test (Army Individual Test Battery, 1944), nonverbal and phonemic fluency (Regard, Strauss, & Knapp, 1982; Thurstone, 1938), the modified Wisconsin Card Sorting Test (Nelson, 1976), digit span and Corsi blocks (Härting, Markowitsch, Neufeld, Calabrese, Deisinger, & Kessler, 2000). All participants were thoroughly medically screened and fulfilled the following inclusion criteria: German as first language; Z scores ≤ −1.96 (2.5th percentile) in no more than 1 of the 11 CERAD-NAB variables; and were in good health, that is, had no current systemic illnesses, no current depression according to the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition [American Psychiatric Association (APA), 1994] as assessed with a standardized questionnaire (Kühner, 1997), no current diseases interfering with the administration of neuropsychological tests (e.g., severe hearing or visual deficits), or diseases of the central nervous system (CNS) at the time of testing. Moreover, participants had no diseases or events during life that could have negatively impacted on CNS activity, and had never been hospitalized for a psychiatric illness. This project was approved by the local Ethics Committee, and written informed consent was obtained from all participants.

Due to resource limitations, only 549 participants were invited for follow-up circa 2 (T_2) and 4 (T_3) years after baseline. T_2 data from participants who remained cognitively healthy at T_3 (*n* = 524) were included in the longitudinal analyses [mean of 2.4 years (*SD* = 0.3) following baseline]. Table 2 lists differences between NC participants with and without follow-up at T_2. The testing protocol at T_2 was identical to that at T_1.

Table 1. Characteristics of healthy aged participants (NC) and patients with probable Alzheimer`s disease (AD) in the cross-sectional sample (T_1)

	NC	AD	Comparisons
N	1,100	352	
Gender (male, female)	690, 410	149, 203	$\chi^2 = 45.5$*
Percentage male	62.7	55.7	
Age ± *SD* (y)	68.7 ± 7.8	75.7 ± 5.9	t = 17.9*
min-max	49 – 92	56 – 89	
Education ± *SD* (y)	12.5 ± 3.0	11.7 ± 3.0	t = 4.2*
min-max	7 – 20	7 – 20	
MMSE ± *SD*	28.9 ± 1.1	26.1 ± 1.6	t = 30.0*
min-max	24 – 30	24 – 30	

Note. y = years. MMSE = Mini-Mental State Examination (Folstein et al., 1975). *$p < .001$.

AD Patients with Very Mild to Mild Dementia

A total of 352 patients from the Memory Clinic at the University Hospital Basel fulfilled inclusion criteria for this study of (1) a diagnosis of probable AD according to the criteria outlined by the National Institute for Neurological and Communicative Disorders and Stroke and the Alzheimer`s Disease and Related Disorders Association (NINCDS-ADRDA; McKhann, Drachman, Folstein, Katzman, Price, & Stadlan, 1984) and DSM-IV (APA, 1994) and (2) an MMSE score ≥ 24/30 (see Table 1). The MMSE inclusion criterion ensured that patients were in the very early stages of the disease. All patients underwent a thorough interdisciplinary examination including a comprehensive neuropsychological assessment (same test battery as for NC sample), a medical examination including neurological status exam, structural brain imaging and laboratory tests (Monsch et al., 1995). We note that, although diagnosing clinicians were not provided with CERAD-NAB total scores, they were presented with individual CERAD-NAB subtest scores together with other neuropsychological test scores, all categorized by cognitive domain. Thus, NC-AD differences in CERAD total scores reported below are most likely artificially inflated. We note that this artificial inflation affects *all* CERAD total scores. As such, the comparison of different CERAD total scores with one another—the primary goal of this study—is not confounded with the inclusion of CERAD data in the diagnostic process. Depressive symptoms were probed with the short version of the Geriatric Depression Scale (GDS; Sheikh & Yasavage, 1986). At T_1, 85% of the AD patients had no depressive symptoms (GDS score 0–4), 11% had mild (GDS score 5–7), 3% moderate (GDS score 8–10), and 2% pronounced (GDS score > 11) depressive symptoms.

Forty-seven percent ($n = 165$) of patients agreed to participate at a follow-up assessment and were tested longitudinally an average of 1.2 years ($SD = 0.3$) after baseline. All patients retained their diagnosis of AD at follow-up, and all retested patient data were included in the longitudinal analyses (see Table 2). The sample of AD patients not available for follow-up included more females and older, less well-educated patients with lower MMSE scores (see Table 2). At T_2, 82% of the patients had no depressive symptoms, 13% had mild, 5% moderate and less than 1% pronounced depressive symptomatology according to the GDS. In the interval between assessments (but not before baseline), $n = 108$ (65%) of the longitudinal patient group were treated with acetylcholine-esterase inhibitors only, $n = 6$ (4%) took part in group memory training only (Ermini-Fünfschilling & Meier, 1995), $n = 21$ (13%) patients received both forms of therapy, and $n = 30$ (18%) patients received no specific dementia therapy.

Material

NC participants and AD patients were administered the German version of the CERAD-NAB (Thalmann et al., 2000) by experienced neurospychologists or trained psychology students. This test battery is comprised of subtests designed to measure those cognitive functions typically affected by AD, that is: "animal fluency" (60 s), a modified version of the "Boston Naming Test" (BNT; maximum score = 15), the MMSE (maximum score = 30), "word list – total" (the sum of words learned after three trials of the 10-word learning list; maximum score = 30), "figures – copy" (maximum score = 11), "word list – delayed recall" (maximum score = 10),

Table 2. Characteristics of healthy aged participants (NC) and patients with probable Alzheimer`s disease (AD) in the longitudinal subsample (T_1 and T_2; follow-up data) and those with no follow-up data

| | Follow-up data | | | No follow-up data | | | |
| | NC | AD | Comparisons | NC | AD | Comparisons | |
	(A)	(B)	(A, B)	(C)	(D)	(A, C)	(B, D)
n	524	165		576	187		
Gender (male, female)	291, 233	81, 84	$\chi^2 = 2.1$	399, 177	68, 119	$\chi^2 = 22.1**$	$\chi^2 = 5.8^\dagger$
Percentage male	55.5	49.1		69.3	36.4		
Age (T_1) ± SD (y)	68.0 ± 7.7	74.8 ± 5.9	t = 12.0**	69.4 ± 7.9	76.5 ± 5.7	t = 2.97*	t = 2.78*
min-max	49 – 88	56 – 87		49 – 92	60 – 89		
Education ± SD (y)	12.4 ± 2.9	12.1 ± 3.2	t = 1.0	12.5 ± 3.0	11.3 ± 2.9	t = 0.8	t = −2.4†
min-max	7 – 20	7 – 20		8 – 20	7 – 20		
MMSE (T_1) ± SD	29.0 ± 1.1	26.4 ± 1.6	t = 18.7**	28.8 ± 1.2	25.9 ± 1.5	t = −2.3†	t = −3.1*
min-max	24 – 30	24 – 30		24 – 30	24 – 30		
MMSE (T_2) ± SD	28.8 ± 1.2	24.5 ± 3.1	t = 17.7**				
min-max	23 – 30	15 – 30					
MMSE ($T_2 - T_1$) ± SD	−0.1 ± 1.3	−1.9 ± 2.7	t = 8.1**				
min-max	−6 – 4	−10 – 5					
Test-Retest-Interval ± SD (y)	2.36 ± 0.3	1.2 ± 0.3	t = 8.9**				
min-max	1.9 – 3.4	0.5 – 3.5					

Note. T_1 = Baseline; T_2 = Follow-up; y = years. MMSE = Mini-Mental State Examination (Folstein et al., 1975). **$p < .001$; *$p < .01$; $^\dagger p < .05$.

"word list – recognition" (maximum score = 100%), and "figures – delayed recall" (maximum score = 11). Three additional variables were created: the number of word responses given during the three word list learning trials and word list delayed recall that were not on the original list was conceptualized as "word list – intrusions"; the proportion of correctly recalled words during the verbal delayed recall compared with verbal learning trial 3 was "word list – savings"; and similarly the proportion of correctly drawn figures during figural delayed free recall compared with the copy condition was "figures – savings". A large, independent study with NC and AD patients demonstrated that these German CERAD-NAB variables have good to excellent discriminative validity (Aebi, 2002).

Statistical Analyses

Global CERAD-NAB scores

Three different global CERAD-NAB scores (all excluding the MMSE) were calculated for the NC and AD participants in the cross-sectional sample at T_1 (cf. Table 1), and for the NC and AD participants in the longitudinal sample at T_1 and T_2 (cf. Table 2): Chandler's total score, a new CERAD-NAB score derived from a principal components analysis of NC data and subsequent logistic regression (PCA-LR score), and a global score based on a logistic regression with jack-knife procedure (LR score). The MMSE was not included in the global score calculation to more directly compare the performance of our total scores with that of Chandler et al. (2005), and because MMSE scores were used as an inclusion criterion for AD patients to participate in this study. To compare the abilities of each measure to discriminate between NC and AD participants, we produced receiver operating characteristics (ROC) curves for each global score and compared the corresponding areas under the curves (see Hanley & McNeil, 1983). For descriptive purposes only, we also report the sensitivity, specificity and correct classification rate (CCR; mean of sensitivity and specificity) of each score.

Chandler score

Chandler CERAD total score was calculated using a procedure identical to Chandler et al. (2005). Raw scores on six German CERAD-NAB variables (i.e., animal fluency with a maximum score of 24 words; modified BNT; word list – total; word list – delayed recall; word list – recognition, subtracting the number of false positives from the number of true positives; figures – copy) were summed, and an age, education and gender corrected regression formula was created with a multiple regression analysis based on the present participants' data.

Principal Components Analysis – Logistic Regression (PCA-LR) score

A principal components analysis (PCA) was conducted to reduce the CERAD-NAB variables to one score while accounting for intercorrelations between variables and minimizing redun-

dancies. CERAD-NAB variables were transformed to achieve normality of standardized residuals (Z scores) in regression models adjusting for covariates (Berres, Zehnder, Bläsi, & Monsch, 2007). The Z scores of the 10 CERAD-NAB variables of the NC group at T_1, excluding MMSE, were used in the PCA. A three-factor PCA solution was selected because three factors achieved eigenvalues greater than 1. Three factor scores were calculated for each participant using coefficient values rounded to one decimal place. These three factor scores were submitted to a logistic regression analysis with backward elimination to produce a global score. The global score was adjusted in a linear model with age, education, and the square of these variables. This analysis resulted in a Z score which was then linearly transformed to a mean of 100 and *SD* of 15.

The PCA-LR score was also calculated for each AD patient. The cutoff score for the PCA-LR scores was determined with a binary logistic regression analysis. A case was classified as "demented" if the predicted probability for dementia was greater than the proportion of true AD patients in the sample. The cutoff value for the binary logistic regression analysis was, therefore, set to a probability of 0.25.

PCA-LR scores for T_2 CERAD-NAB performance were calculated using a procedure identical to the one described above.

Logistic regression (LR) score

Stepwise logistic regression analyses with backward elimination (exclusion criterion: $p = .10$; inclusion criterion: $p = .05$; cutoff value = 0.25) comparing NC participants to AD patients were performed on all 10 z-transformed baseline CERAD-NAB scores (excluding MMSE). Logistic regression estimates the optimal weighting of each subtest score, as determined by the best fit of predicted probabilities to observed outcomes. Thus, it tends to give high weights to variables with high measurement precision, similar to methods which directly weight by precision (see, e.g., Wouters, van Gool, Schmand, & Lindeboomm, 2008), but has the added advantage of accounting for correlations between variables.

Cross-validation

To determine the diagnostic accuracies of each baseline CERAD-NAB total score, we first randomly split the NC and AD baseline samples (group A and group B) for cross-validation analyses (see Table 3 for demographic information). Thus, total scores were derived from group A and applied to group B, and in a second step, scores were derived from group B and applied to group A. The diagnostic discriminability of each set of scores was quantified with sensitivity, specificity, CCR, and AUC.

Entire Sample

To foreshadow the cross-validation results, each set of analyses resulted in comparable diagnostic discriminabilities (see Table 4), indicating a high stability of the generated total scores. To produce a single set of total score formulae, we next calculated CERAD total scores on the entire participant

Table 3. Characteristics of healthy aged participants (NC) and patients with probable Alzheimer`s disease (AD), randomly assigned to groups A and B for cross-validation

	Group A	Group B	Comparisons
NC			
n	550	550	
Gender (male, female)	342, 208	348, 202	$\chi^2 = .140$
Percentage male	62.2	63.3	
Age ± *SD* (y)	68.5 ± 8.0	68.9 ± 7.7	$t = -.747$
Education ± *SD* (y)	12.5 ± 3.1	12.4 ± 2.9	$t = .580$
MMSE ± *SD*	28.9 ± 1.1	28.8 ± 1.2	$t = .817$
AD			
n	176	176	
Gender (male, female)	75, 101	74, 102	$\chi^2 = .012$
Percentage male	42.6	42.0	
Age ± *SD* (y)	75.4 ± 6.0	76.0 ± 5.8	$t = -.897$
Education ± *SD* (y)	11.6 ± 3.0	11.8 ± 3.2	$t = -.645$
MMSE ± *SD*	26.2 ± 1.6	26.1 ± 1.6	$t = .541$

Note. y = years. MMSE = Mini-Mental State Examination (Folstein et al., 1975).

population. For the LR score based on the entire sample, the stability of the regression model was tested with a cross-validation via a jackknife procedure ("leave one out" method). This procedure was used to calculate LR scores at T_1 and T_2. We report the results (sensitivity, specificity, CCR, AUC) of these analyses after the jackknife cross-validation.

Global Change Scores

To determine the extent to which the global scores discriminated between the longitudinal performance of NC and AD participants, difference scores were calculated by subtracting each score at T_1 from the corresponding score at T_2. Because difference scores based on Z scores are also influenced by demographic variables, difference scores were corrected for the effects of age, education, gender, baseline performance, and all possible interactions between these variables (Zehnder et al., 2007). Forty-five regression models were fitted for each global score. The most accurate model, that is, the model with the smallest standard deviation of the predicted residuals, was selected based on the results of the Predicted Residual Sum of Squares statistics (Berres et al., 2007).

Combination of Cross-sectional and Longitudinal Data

To determine whether the correct classification rate based on combined information from (a) T_1 and T_2 and (b) T_1 and

Table 4. Results of the double cross-validation of randomly split groups

	Training data			Validation data		
	Group A			Group B		
	Chandler score	PCA-LR score	LR score	Chandler score	PCA-LR score	LR score
Sensitivity	90.9	90.9	93.2	93.2	93.2	90.9
Specificity	91.5	90.2	91.3	90.0	90.4	90.7
CCR	91.2	90.6	92.3	91.6	91.8	90.8
AUC	.967	.972	.980	.970	.968	.973
	Group B			Group A		
	Chandler score	PCA-LR score	LR score	Chandler score	PCA-LR score	LR score
Sensitivity	91.5	92.0	92.0	88.6	89.8	88.6
Specificity	91.3	90.5	92.4	92.9	90.4	92.7
CCR	91.4	91.3	92.2	90.8	90.1	90.7
AUC	.969	.966	.978	.965	.969	.972

Note. PCA-LR = logistic regression on principal components analysis scores; LR = logistic regression on demographically corrected (CERAD-NAB) variables; CCR = correct classification rate; AUC = area under the curve. The sensitivity, specificity, CCR, and AUC in the validation sample are based on application of score formulae derived from the training data.

(T_2–T_1) is superior to the diagnostic accuracy of T_1 information alone, linear combinations of these data were calculated for the Chandler and PCA-LR global scores. These values were compared with scores from the logistic regression with backward elimination (exclusion criterion: $p = .10$, inclusion criterion: $p = .05$) based on all 10 z-transformed CERAD-NAB variables at timepoints T_1 and T_2.

Statements of significance refer to comparison-wise error rates and should be interpreted in a descriptive sense.

RESULTS

Cross-validation

The results of the cross-validation analyses are shown in Table 4. Because the diagnostic discriminabilities of scores in the validation and cross-validation samples were comparable, and to generate a single set of total score formulae, all ensuing analyses were conducted on the entire dataset.

Cross-sectional Sample (cf. Table 1)

The mean uncorrected baseline Chandler total score of the NC group was 81.3 ($SD = 8.4$; range = 53–99), and of the AD patients was 56.3 ($SD = 10.0$; range = 33–83). The mean adjusted Chandler total score for the NC group was 95.4 ($SD = 7.3$; range = 63.5–114.1), higher than that of the AD patients (mean = 73.0; $SD = 9.4$; range = 47.1–100.9) ($t[494.3] = 40.9$, $p < .001$, two-tailed), and the cutoff score was 85.11. The formula for the demographic correction was: Chandler raw total score – (- 0.391*age + 0.886*education + 4.447*gender), where gender is coded as man = 0, woman = 1. This adjusted Chandler total score was highly correlated with the adjusted American version ($r = 0.997$).

The PCA-LR global CERAD-NAB score was based on a PCA and logistic regression. Table 5 contains the factor loadings resulting from the PCA of all CERAD-NAB subtest performance of the NC group at T_1. Factor 1 represents "verbal memory," factor 2 "nonverbal memory," and factor 3 "non-memory functions." The logistic regression of the raw PCA factor loadings resulted in the following PCA-LR raw score: = 4.355 + 2.23*factor1 + 0.466*factor3. The PCA-LR raw scores were transformed with t_PCA-LR score = sign(PCA-LR raw score) * abs(PCA-LR raw score)$^{1.15}$, with sign(x) = 1 for x $\geq$ 0 and = −1 for x < 0 in preparation for the adjustments for linear and quadratic effects of the demographic variables. The corrected PCA-LR global Z score was: Z_PCA-LR score = [t_PCA-LR score – (4.779 + 0.00701*age + 0.05718*education - 0.00246*(age - 68.68)^2 - 0.02592*(education - 12.47)^2)] / 3.204. The NC data were then linearly transformed to a mean of 100 and SD of 15 (range = 51.65–144.04). The mean PCA-LR score for the AD patients (mean = 60.34; $SD = 15.75$; range = 17.2–109.85) was lower than that of the NC participants ($t[1450] = 42.7$; $p < .001$, two-tailed), and the resulting cutoff score was 80.13.

The numerical values for the Chandler and PCA-LR scores cannot be directly compared because they underwent different transformations.

The logistic regression analysis resulted in a model containing all 10 z-transformed CERAD-NAB variables which provided the best possible discrimination between groups: 4.961 + (.704*animal fluency) + (.199*BNT) + (.719*word list – total) + (.293* word list – intrusions) + (.459*figures – copy) + (1.626 word list – delayed recall) – (.408*word list – savings) + (.295*word list – recognition) – (.679*figures – delayed recall) + (1.437*figures – savings).

To estimate the ability of the global scores to discriminate between NC and AD participants, ROC curves were generated for each global score. The sensitivity and specificity, as well as the CCR of each global score were calculated from these ROC curves (see Table 6). Comparisons of the areas under the ROC curve revealed that the LR score showed trends to discriminate better than the PCA-LR score ($Z = 1.96$; $p = .05$) and the Chandler score ($Z = 1.80$; $p = .07$). The Chandler and PCA-LR scores performed comparably ($Z = 0.77$; $p = .44$). The distributions of the Chandler and PCA-LR scores for NC and AD participants with corresponding

Table 5. Factor loadings of the German CERAD-NAB variables resulting from a principal component analysis (PCA; 3-factor solution) of healthy aged participants' performance ($N = 1,100$)

CERAD variable	Factor 1	Factor 2	Factor 3
Animal Fluency	0.115	0.020	0.487
Boston Naming Test	0.109	0.078	0.344
Word List – Total	0.225	−0.110	0.112
Word List – Intrusions[a]	0.129	−0.130	−0.261
Figures – Copy	0.054	0.102	0.543
Word List – Delayed Recall	0.290	−0.219	0.036
Word List – Savings[b]	0.228	−0.209	−0.156
Word List – Recognition	0.202	−0.169	0.026
Figures – Delayed Recall	0.191	0.475	−0.060
Figures – Savings[c]	0.178	0.450	−0.289

Note. CERAD-NAB = Consortium to Establish a Registry on Alzheimer's Disease-Neuropsychological Assessment Battery.
[a]Number of words provided during the three learning trials and the Word List – Delayed Recall that were not on the Word List.
[b]Proportion of correctly recalled words during the Word List - Delayed Recall compared to the Word List learning trial 3.
[c]Proportion of correctly drawn figures during Figures - Delayed Recall compared to Figures - Copy.

Table 6. Comparisons of the discriminatory diagnostic characteristics of the Chandler score, PCA-LR score, and LR score in the cross-sectional (T_1) and longitudinal (T_1 and T_2) samples

	Chandler score	PCA-LR score	LR score
Cross-sectional sample (T_1)			
Sensitivity	91.2	91.2	91.2
Specificity	91.6	90.4	92.0
CCR	91.4	90.8	91.6
AUC	.968	.968*	.976
Longitudinal sample (T_1)			
Sensitivity	89.1	90.3	88.5
Specificity	92.0	89.3	93.1
CCR	90.6	89.8	90.8
AUC	.960	.962	.969
Longitudinal sample (T_2)			
Sensitivity	88.5	92.7	90.3
Specificity	93.1	91.6	93.3
CCR	90.8	92.2	91.8
AUC	.967	.971	.974

Note. PCA-LR = logistic regression on principal components analysis scores; T_1 = baseline; T_2 = follow-up; CCR = correct classification rate; AUC = area under the curve; LR = logistic regression on demographically corrected (CERAD-NAB) variables, results after jack-knife validation.
*Score differs from LR score ($p = .05$).

cut-off scores are shown in Figure 1a,b. These figures illustrate the excellent diagnostic discriminabilities of the Chandler and PCA-LR scores, as well as the number of false classifications associated with other, suboptimal, cut-off scores.

Longitudinal Subsample (cf. Table 2)

Baseline

At T_1, the longitudinal subsample of NC participants had higher Chandler scores (mean = 95.42; SD = 6.95; range = 73.61–113.7) and PCA-LR scores (mean = 99.95; SD = 14.4; range = 51.65–134.74) than the AD patients (Chandler score: mean = 74.27; SD = 9.65; range = 47.11–100.91; $t[220.1] = 26.1$; $p < .001$, two-tailed; PCA-LR score: mean = 63.21; SD = 15.81; range = 28.19–109.85; $t[687] = 27.9$; $p < .001$, two-tailed). The area under the ROC curve of the Chandler score did not significantly differ from the area under the ROC curve of the PCA-LR score ($Z = 0.23$; $p = .82$) or the LR score ($Z = 1.18$; $p = .24$), and the areas under the ROC curves of the PCA-LR and LR scores also did not significantly differ ($Z = 1.07$; $p = .28$) (see Table 6).

Follow-up

At T_2, the NC participants still had a greater Chandler score (mean = 98.01, SD = 6.31) and PCA-LR score (mean = 105.26, SD = 13.76) compared with the AD patients (Chandler score: mean = 73.09; SD = 11.64; $t[195.3] = 26.31$; $p < .001$; PCA-LR score: mean = 60.45; SD = 18.75; $t[222.4] = 28.39$; $p < .001$), as expected.

The LR analysis resulted in a model containing five z-transformed CERAD-NAB variables which provided the best possible discrimination between groups: 3.767 + (.908*

animal fluency) – (1.219*word list – total) + (0.603*word list – delayed recall) + (.581*word list – recognition) + (.731*figures – delayed recall).

The Chandler and PCA-LR scores demonstrated similar discriminatory abilities ($Z = 0.44$; $p = .66$), as did the LR and Chandler scores ($Z = 1.02$; $p = .31$) as well as LR and PCA-LR scores ($Z = 0.57$; $p = .57$). In this subsample, the optimal cut-off score for the Chandler score was 85.89 at T_1 and 88.1 at T_2, for the PCA-LR score 81.3 at T_1 and 84.82 at T_2 (see Table 6).

Longitudinal analyses

Difference scores (T_2–T_1) were calculated for the Chandler and PCA-LR and standardized to a mean of 0 and a SD of 1 in the NC sample. Both difference scores for the AD patients were significantly different from zero (Chandler score: mean = −2.62; SD = 1.63; range = −6.33–2.16; $t[203.7] = 19.42$; $p < .001$; PCA-LR score: mean = −2.19; SD = 1.21; range = −5.07–1.63; $t[238] = 21.06$; $p < .001$), indicating a significant decline in CERAD-NAB performance over time. The sensitivities, specificities and CCRs for the Chandler difference score were 80.6, 87.8 and 84.2 (AUC = 0.91), respectively, and the corresponding values for the PCA-LR difference score were 85.5, 85.1 and 85.3 (AUC = 0.922), respectively. The LR difference score based on the z-transformed differences in Z score between T_2 and T_1 and following a jackknife procedure demonstrated a sensitivity of 85.5, a specificity of 90.3 and a CCR of 87.9 (AUC = 0.934). All three measures did not significantly differ with respect to their abilities to discriminate NC from AD participants' longitudinal performance (AUC comparisons; Chandler *vs.* PCA-LR score: $Z = 0.79$; Chandler *vs.* LR: $Z = 1.60$; PCA-LR *vs.* LR: $Z = 0.95$).
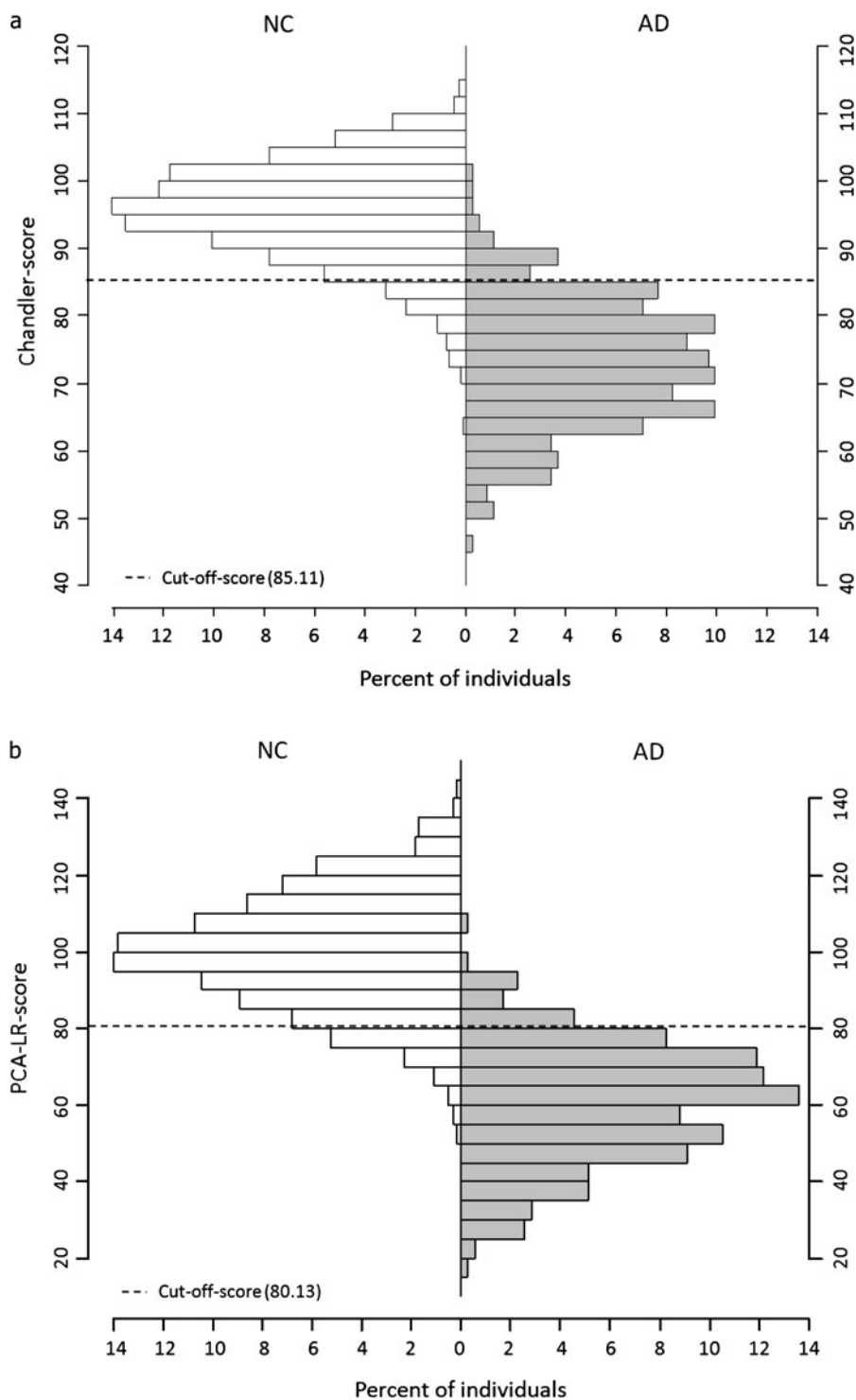
Fig. 1. Distribution of (a) Chandler and (b) logistic regression on principal components analysis (PCA-LR) scores in healthy elderly individuals (NC) and patients with Alzheimer's disease (AD) at baseline (T_1). Each optimal cutoff score classifies individuals with higher scores as "healthy" and those with lower scores as "demented".

Combination of cross-sectional and longitudinal information

The results of the binary logistic regression for both combinations [T_1 and T_2 as well as T_1 and difference scores ($T_2 - T_1$)] for each global CERAD-NAB score are shown in Table 7.

Both sets of scores were comparable in their ability to discriminate NC from AD participants according to comparisons of the corresponding areas under the curves (all $Z <$ 1.49). While both combinations of PCA-LR scores demonstrated a higher diagnostic accuracy than PCA-LR scores based on T_1 alone (n.b. areas under the curve identical for

Table 7. Discriminatory diagnostic characteristics of the Chandler score, PCA-LR score, and LR score for the combined cross-sectional and longitudinal data

	Chandler score T_1 & T_2	Chandler score T_1 & difference	PCA-LR score T_1 & T_2	PCA-LR score T_1 & difference	LR score z-scores T_1 & T_2
Sensitivity	89.1	89.1	90.3	90.9	94.5
Specificity	93.1	93.3	92.0	91.8	93.5
CCR	91.1	91.2	91.2	91.4	94.0
AUC	.974	.974	.978	.978	.979

Note. PCA-LR = logistic regression on principal components analysis scores; T_1 = baseline; T_2 = follow-up; difference = adjusted difference score ($T_2 - T_1$); CCR = correct classification rate; AUC = area under the curve; LR = logistic regression on demographically corrected (CERAD-NAB) variables, results after jackknife validation.

both PCA-LR combination scores; $Z = 2.27$, $p = .02$), there was only a trend for the Chandler combination scores to outperform the Chandler score for T_1 (n.b. areas under the curve identical for both Chandler combination scores; $Z = 1.88$, $p = .06$), and the already excellent discriminatory power of the LR score at baseline did not improve when information from T_1 and T_2 were combined ($Z = 1.51$, $p = 0.13$).

DISCUSSION

All global CERAD-NAB scores examined here demonstrated a comparably good ability to correctly classify NC participants and very early AD patients. Moreover, the consistency of the total score performance measures in the cross-validation analyses indicate that these measures are highly stable. The logistic regression represents an ideal method for diagnostic discrimination at the group level, and the supplementary jackknife procedure optimizes logistic regression scores for future samples. The present analyses, in which global scores were stringently compared with the logistic regression with jackknife procedure, showed that the Chandler score and the PCA-LR score both have excellent diagnostic discriminability for NC and early AD (we should note that the numerical values for the Chandler and PCA-LR scores cannot be directly compared because of the different methods of derivation and scalings). However, the relative ease of calculating the Chandler et al. score makes it a practical choice for the assessment of early Alzheimer's disease patients with the German CERAD-NAB.

The CCRs were most likely affected by several competing factors. With respect to the change scores, stringent statistical analyses were used because the test–retest interval of the NC participants was nearly double that of the AD patients, who most likely would have shown an even more pronounced cognitive decline with an equally long test–retest interval. Thus, the relatively high CCRs between 84.2% and 87.9% most likely underestimate the potential of these scores to correctly classify individuals on the basis of their change in cognitive performance over time and thus also the constructed combination scores. Similar to our global scores, Zehnder et al. (2007) reported that individual CERAD-NAB subtest baseline and change scores both showed excellent diagnostic discriminability for NC and AD in groups who had the same test–retest intervals as the present study. More-

over, 78% of AD patients with follow-up testing had received acetylcholine-esterase inhibitor treatment between baseline and follow-up, which may also have weakened differences between the CERAD total scores of NC and AD participants. It would be interesting for further studies to quantify the potential effects of such treatments on longitudinal CERAD-NAB total score performance. AD patients available for follow-up also tended to be younger, better educated, and have higher MMSE scores than AD patients who were not followed-up, factors presumably associated with higher CERAD-NAB scores and a decreased CCR. However, an additional factor may have led to an overestimation of the CCR: CERAD-NAB scores were available to clinicians during the diagnostic process. This partial circularity, which affected all total scores, most likely artificially increased their CCRs.

The combined baseline and longitudinal PCA-LR scores provided significantly greater diagnostic discriminability compared with baseline PCA-LR data alone, whereby combined Chandler score information showed a trend to outperform corresponding baseline data. These findings suggest that the consideration of information from two testing sessions can be diagnostically relevant, especially in the early stages of a dementing illness or in cases where uncertainties surround the initial diagnosis (e.g., initially good test performance combined with caregiver report or clinical signs of impairments in activities of daily living). Indeed, the results of a recent study (Rosetti, Cullum, Hynan, & Lacritz, 2010) support the utility of the CERAD total score to measure the progression of global neuropsychological impairment in AD.

Because also the Chandler total score was developed from NC data, it may additionally prove useful in discriminating NC individuals from those with other forms of dementia. Aebi (2002) demonstrated that 7 of the 10 CERAD-NAB variables discriminated NC from patients with AD, vascular dementia and mixed dementia with an accuracy between 81 and 86%. Moreover, a total score based on a newly developed, extended version of the German CERAD-NAB ("plus") battery which includes phonemic fluency (S-words; Thurstone, 1938) and the TMT (Army Individual Test Battery, 1944) may prove especially useful in discriminating NC individuals from patients with subcortical forms of dementia.

The higher diagnostic discriminability (NC *vs.* AD) of the total score in Chandler et al.'s (2005) population (sensitivity 93.7% / specificity 92.6% *vs.* 92.0% / 89.1% in the present sample) may reflect differences in demographic correction formulae and the more advanced stage of AD in Chandler et al.'s population. When the original American demographic correction formula was applied to our sample, mean corrected CERAD total scores of our AD patients (71.6; *SD* = 9.3) were more comparable to those of Chandler et al.'s MCI sample (76.9; *SD* = 8.9) than their AD sample (60.2; *SD* = 11.9). These findings highlight the challenge of differentiating patients with (amnestic) MCI from those in a very early stage of dementia (Winblad et al., 2004). This differentiation will depend in part on assessing complex instrumental activities in daily living as precisely as possible, although it remains unclear what these activities are and how they can be discriminated from activities of daily living. This is an important topic for further investigation, especially because patients with a minimal degree of symptomatology would be expected to profit from a sufficiently safe and tolerable treatment to delay the progression of the disease, for example, immunization therapy (Grimmer, Perneczky, & Kurz, 2008). For this reason, and in light of recent developments in amyloid ß research, the early diagnosis of AD is of central importance (Forsberg et al., 2008; Pike et al., 2007). The CERAD neuropsychological assessment battery appears to be an excellent tool for this purpose.

Complex cognitive profiles composed of a large number of individual test scores can be a cumbersome form of communicating an individual's cognitive status, especially in intervention or longitudinal studies where individual scores can show independent and in some cases divergent patterns over time. In many cases, instruments such as the ADAS-Cog have been used in addition to the MMSE (Brodaty, Corey-Bloom, Potocnik, Truyen, Gold, & Damaraju, 2005; Rogers, Farlow, Doody, Mohs, & Friedhoff, 1998; Seltzer et al., 2004; Tariot, Solomon, Morris, Kershaw, Lilienfeld, & Ding, 2000). The CERAD-NAB has the advantage over the ADAS-Cog and MMSE of measuring both delayed recall and recognition from episodic memory, functions critically impaired in AD. The CERAD total score based on demographically corrected raw scores represents a practical alternative tool to assess global cognitive functioning. During repeated testing with the validated CERAD-NAB, this score allows for easily communicable conclusions to be made about the course of cognitive functioning and, therefore, appears especially suited to intervention studies, for example, to assess the potential modification of disease progression with different kinds of pharmacotherapy.

For the German version of the CERAD-NAB, Chandler et al.'s (2005) method of calculating a total score adjusted for the influences of demographic variables can be recommended. This score is much simpler to calculate than the global CERAD-NAB score based on a principal component analysis (i.e., PCA-LR score) and provides an effective global measure of cognitive functioning.

ACKNOWLEDGMENTS

REFERENCES

Aebi, C. (2002). *Validierung der neuropsychologischen Testbatterie CERAD-NP: eine Multi-Center Studie [Validation of the CERAD neuropsychological assessment battery: A multi centre study]*. Unpublished doctoral dissertation, University of Basel, Basel, Switzerland.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental Disorders* (4th ed.). Washington, DC: American Psychiatric Association Press.

Army Individual Test Battery. (1944). *Manual of directions and scoring*. Washington, DC: War Department, Adjutant General's Office.

Berres, M., Zehnder, A., Bläsi, S., & Monsch, A.U. (2007). Evaluation of diagnostic scores with adjustment for covariates. *Statistics in Medicine*, *27*, 1777–1790.

Brodaty, H., Corey-Bloom, J., Potocnik, F.C., Truyen, L., Gold, M., & Damaraju, C.R. (2005). Galantamine prolonged-release formulation in the treatment of mild to moderate Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *20*, 120–132.

Chandler, M.J., Lacritz, L.H., Hynan, L.S., Barnard, H.D., Allen, G., Deschner, M., et al. (2005). A total score for the CERAD neuropsychological battery. *Neurology*, *65*, 102–106.

Ermini-Fünfschilling, D., & Meier, D. (1995). Gedächtnistraining: Wichtiger Bestandteil der Milieutherapie bei seniler Demenz [Memory training: An important constituent of milieu therapy in senile dementia]. *Zeitschrift für Gerontologie und Geriatrie*, *28*, 190–194.

Fillenbaum, G.G., Unverzagt, F.W., Ganguli, M., Welsh-Bohmer, K.A., & Heyman, A. (2002). The CERAD Neuropsychological Battery: Performance of representative community and tertiary care samples of African-American and European-American elderly. In F.R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment*. Lisse, Netherlands: Swets & Zeitlinger.

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*, 189–198.

Forsberg, A., Engler, H., Almkvist, O., Blomquist, G., Hagman, G., Wall, A., et al. (2008). PET imaging of amyloid deposition in patients with mild cognitive impairment. *Neurobiology of Aging*, *29*, 1456–1465.

Grimmer, T., Perneczky, R., & Kurz, A. (2008). Aktueller Stand der Immuntherapie bei Alzheimer-Krankheit [Current immune therapy for Alzheimer's disease]. *Nervenarzt*, *79*, 832–835.

Hanley, J.A., & McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*, 839–843.

Härting, C., Markowitsch, H.J., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (Eds.). (2000). *Wechsler Gedächtnistest – revidierte Fassung [Wechsler Memory Scale – revised edition]*. Bern, Switzerland: Huber.

Kühner, C. (1997). *Fragebogen zur Depressionsdiagnostik nach DSM-IV (FDD DSM-IV) [German Version of the Inventory to Diagnose Depression according to DSM-IV]*. Göttingen, Germany: Hogrefe.

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*, 939–944.

Monsch, A.U., Foldi, N.S., Ermini-Fünfschilling, D.E., Berres, M., Taylor, K.I., Seifritz, E., et al. (1995). Improving the diagnostic accuracy of the Mini-Mental State Examination. *Acta Neurologica Scandinavica*, *92*, 145–150.

Monsch, A.U., Thalmann, B., Schneitter, M., Bernasconi, F., Aebi, C., Camachova Davet, Z., et al. (2000). The Basel Study on the Elderly's search for preclinical cognitive markers of Alzheimers's disease. *Neurobiology of Aging*, *21*, 31.

Morris, J.C., Edland, S., Clark, C., Galasko, D., Koss, E., Mohs, R., et al. (1993). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part IV. Ratings of cognitive change in the longitudinal assessment of probable Alzheimer's disease. *Neurology*, *43*, 2457–2465.

Morris, J.C., Heyman, A., Mohs, R.C., Hughes, J.P., van Belle, G., Fillenbaum, G., et al. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, *39*, 1159–1165.

Morris, J.C., Mohs, R.C., Rogers, H., Fillenbaum, G., & Heyman, A. (1988). Consortium to Establish a Registry for Alzheimer's Disease (CERAD) clinical and neuropsychological assessment of Alzheimer's disease. *Psychopharmacology Bulletin*, *24*, 641–652.

Nelson, H.E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*, 313–324.

Panisset, M., Roudier, M., Saxton, J., & Boller, F. (1994). Severe impairment battery: A neuropsychological test for severely demented patients. *Archives of Neurology*, *51*, 41–45.

Pike, K.E., Savage, G., Villemagne, V.L., Ng, S., Moss, S.A., Maruff, P., et al. (2007). Beta-amyloid imaging and memory in non-demented individuals: Evidence for preclinical Alzheimer's disease. *Brain*, *130*, 2837–2844.

Regard, M., Strauss, E., & Knapp, P. (1982). Children's production on verbal and non-verbal fluency tasks. *Perceptual and Motor Skills*, *55*, 839–844.

Rogers, S.L., Farlow, M.R., Doody, R.S., Mohs, R., & Friedhoff, L.T. (1998). A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Donepezil Study Group. *Neurology*, *50*, 136–145.

Rosen, W.G., Mohs, R.C., & Davis, K.L. (1984). A new rating scale for Alzheimer's disease. *American Journal of Psychiatry*, *141*, 1356–1364.

Rosetti, H.C., Cullum, C.M., Hynan, L.S., & Lacritz, L.H. (2010). The CERAD neuropsychologic battery total score and the progression of Alzheimer disease. *Alzheimer Disease and Associated Disorders*, *24*, 138–142.

Seltzer, B., Zolnouni, P., Nunez, M., Goldman, R., Kumar, D., Ieni, J., et al. (2004). Efficacy of donepezil in early-stage Alzheimer's disease: A randomized placebo controlled trial. *Archives of Neurology*, *61*, 1852–1856.

Sheikh, J.I., & Yasavage, J.A. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. In T.L. Brink (Ed.), *Clinical gerontology: A guide to assessment and intervention* (pp. 165–174). New York: The Haworth Press.

Tariot, P.N., Solomon, P.R., Morris, J.C., Kershaw, P., Lilienfeld, S., & Ding, C. (2000). A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology*, *54*, 2269–2276.

Thalmann, B., Monsch, A.U., Schneitter, M., Bernasconi, F., Aebi, C., Camachova Davet, Z., et al. (2000). The CERAD neuropsychological assessment battery (CERAD-NAB) – A minimal dataset as a common tool for German-speaking Europe. *Neurobiology of Aging*, *21*, 30.

Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Welsh, K.A., Butters, N., Hughes, J., Mohs, R., & Heyman, A. (1991). Detection of abnormal memory in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Archives of Neurology*, *48*, 278–281.

Welsh, K.A., Butters, N., Hughes, J., Mohs, R., & Heyman, A. (1992). Detection and staging of dementia in Alzheimer's disease: Use of the neuropsychological measures developed for the Consortium to Establish a Registry for Alzheimer's disease (CERAD). *Archives of Neurology*, *49*, 448–452.

Welsh, K.A., Butters, N., Mohs, R.C., Beekly, D., Edland, S., Fillenbaum, G., et al. (1994). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology*, *44*, 609–614.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., et al. (2004). Mild cognitive impairment – beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, *256*, 240–246.

Wouters, H., van Gool, W.A., Schmand, B., & Lindeboom, R. (2008). Revising the ADAS-cog for a more accurate assessment of cognitive impairment. *Alzheimer Disease and Associated Disorders*, *22*, 236–244.

Zehnder, A.E., Bläsi, S., Berres, M., Spiegel, R., & Monsch, A.U. (2007). Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? *American Journal of Alzheimer's Disease and Other Dementias*, *22*, 416–426.