Lack of Practice Effects on Neuropsychological Tests as Early Cognitive Markers of Alzheimer Disease?

Antoinette E. Zehnder, Stafan Bläsi, Manfred Berres, Rene Spiegel and Andreas U. Monsch

Am J Alzheimers Dis Other Demen 2007; 22; 416
DOI: 10.1177/1533317507302448

American Journal of Alzheimer's Disease & Other Dementias®
Volume 22 Number 5
October/November 2007 416-426

10.1177/1533317507302448
http://ajadd.sagepub.com
hosted at
http://online.sagepub.com

Lack of Practice Effects on Neuropsychological Tests as Early Cognitive Markers of Alzheimer Disease?

Antoinette E. Zehnder, PhD candidate, Stafan Bläsi, MA, Manfred Berres, PhD, Rene Spiegel, PhD, and Andreas U. Monsch, PhD

Reliable assessment of change from previous cognitive functioning is a prerequisite for determining the possible presence of neurodegenerative diseases such as Alzheimer's disease (AD). We investigated whether standardized change scores on the German version of the Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery (CERAD-NAB) could be used for early diagnosis of AD and whether change scores on the CERAD-NAB are superior in this respect to scores recorded on 1 occasion only. Three hundred seventy-four normal control subjects were assessed twice. Data from 95 patients with mostly mild probable AD were collected at their first entry to a memory clinic and an average of 1.1 ± 0.24 years later. It is concluded that repeated testing with the CERAD-NAB does not generally add to improved diagnostic accuracy for mild and very mild AD and cannot, therefore, be recommended as a routine clinical procedure.

Keywords: neuropsychological change; early dementia; longitudinal assessment; change score

In aged persons with observed or subjectively noticed mental decline, a reliable diagnosis of dementia should be made as early as possible in order (1) to detect reversible forms of dementia and institute appropriate treatment; (2) to enable patients to benefit from treatment that could delay the progression or even the onset of the manifest disease; (3) to delay placement in an institution with its undesirable emotional, behavioral, and economic consequences; and (4) to allow patients sufficient time to make mentally competent arrangements, such as his or her last will and/or advance directives for the family to adjust to the situation. Among the different causes of dementia, Alzheimer's disease (AD) is the most common, accounting for about three quarters of all cases.[1] The neurodegenerative process underlying AD is thought to begin years or even decades before clinical symptoms appear.[2] Studies indicate that the preclinical period of AD is characterized by an early onset (6 to 3 years) of some cognitive impairment, followed by relative stability until a few years (2 to 3 years) before diagnosis, when precipitous cognitive decline occurs[3]—meaning that patients can be in a more or less symptom-free preclinical stage of AD for considerable time in which it might be possible to identify them.

It has been suggested that neuropsychological tests such as measures of episodic memory (verbal as well as nonverbal),[4] story recall tasks (immediate and delayed), naming tasks (Boston Naming Test),[5] and tests of executive functioning[6] are particularly suited as early and possibly preclinical markers of AD; however, even the use of age- and education-adjusted norms cannot account for individual slopes of decline if tests are performed on 1 occasion only. Subjects with a low premorbid level of functioning are likely to fall below cutoff values more quickly, whereas patients with a high premorbid level of functioning are more likely to remain above cutoff scores for some

From the Memory Clinic-Neuropsychology Center, University Hospital, Basel, Switzerland.

Address correspondence to: Andreas U. Monsch, PhD, Memory Clinic-Neuropsychology Center, University Hospital, Schanzenstrasse 55, 4031 Basel, Switzerland; e-mail: andreas.monsch@unibas.ch.

time. To remedy this situation, it has been proposed[7,8] that cognitive change, as assessed in a longitudinal approach, may be more indicative of the development of dementia than cross-sectional assessments only.

Before statements about significant cognitive decline in potential patients can be made, the normal degree of cognitive change in a healthy aged population must be determined. The observation that healthy aged subjects very often showed superior performance at repeated neuropsychological testing drew our attention to the phenomenon of practice effects on neuropsychological tests. Practice effects at repeated testing have been reported to occur among brain-injured patients,[9] in children,[10] and in adults across different age categories.[11] For older subjects, they are usually greater on simple tasks and smaller on complex tasks,[12] which may be due to age-related decline in adaptability.[11] They vary in magnitude between different tasks, being especially observed on tests that are timed,[13] those that involve psychomotor processing and[14] learning,[15] or those that involve learning of specific rules or problem-solving strategies.[16]

Practice effects may be due to a number of factors, ranging from increased familiarity with the test-taking situation or procedures ("test sophistication")[17] to remembering specific test items. Depending on the situation, practice effects are seen as a confounding variable because better familiarity or remembering test items may mask cognitive decline, resulting in apparent stability or only slight deterioration in cognitive test performance when in fact significant decline has occurred. As a consequence, practice effects need to be accounted for in the context of repetitive testing, and there have been different approaches to this problem.[12,18] In contrast, the aim of this study was not to adjust the empirical findings for practice effects, but to investigate changes of performance on repeated testing as potential early markers of AD. Specifically, we aimed to standardize changes of performance seen in healthy aged subjects (referred to as practice effects) on different cognitive tasks and to compare them with changes seen in patients with mild AD. Stable or improved performance (practice effects) may then reflect the absence of cognitive deterioration, whereas a lack of enhanced performance in some tasks after repeated testing may be an indicator of pathology, as AD patients do not recall the circumstances and any details of the test situation, the test materials, procedures, approaches, and so forth. This failure of recall (episodic memory) is assumed to prevent them from improving their actual performance when being tested for a second or third time. Based on these considerations, we hypothesized that change scores on neuropsychological tests might be superior as early clinical or even preclinical markers of AD than the same measures taken on 1 occasion only.

Methods

Healthy Aged Subjects

The healthy participants of this study are a subsample of the Basel Study on the Elderly (BASEL)[19] cohort. BASEL is a follow-up project to the Basel Longitudinal Studies,[20] which was initiated in 1997; its aims are to identify in-life biological and preclinical cognitive markers of AD. The project was approved by the local ethics committee, and written informed consent was obtained from all participants. Baseline assessments (T_0) were performed between 1997 and 2001. All participants underwent a thorough clinical examination, including a detailed medical history questionnaire, neuropsychological evaluation with the German version of the CERAD-NAB,[21] the Clock Drawing Test,[22] and an assessment of depressive symptoms (*Fragebogen zur Depressionsdiagnostik Nach DSM-IV*).[23] Blood was drawn to determine the apolipoprotein E genotype, and DNA samples were frozen for later genetic analyses.

At an average of 2.4 ± 0.28 years (range, 1.9-3.4 years) later, participants underwent a follow-up examination (T_1) using the same assessment instruments. For the current analyses, participants who had completed at least 2 assessments (ie, T_0 and T_1) and who fulfilled the following criteria at both time points were selected:

1. z scores ≤ −1.96 (ie, abnormally low values; 2.5th percentile) in not more than 1 of the 11 standard CERAD-NAB variables
2. Speak German as their first language
3. In good health condition, that is, had no current systemic illnesses, no diseases interfering with the administration of neuropsychological tests (eg, severe hearing or visual deficits), no psychiatric problems, no diseases of the central nervous system; and no diseases or events during life that could have had a negative impact on central nervous system activity
4. Did not suffer from depression according to the *Diagnostic and Statistical Manual of Mental Disorders*[24] criteria, as assessed with a standardized questionnaire.[23]

Table 1. Baseline Characteristics of Normal Control Subjects and Patients with Probable Alzheimer's Disease

	Normal Control Subjects (Baseline)	Alzheimer's Disease Patients (Baseline)	Comparisons
N	374	95	
Gender (male, female)	246, 128	38; 57	$\chi^2_{(1)} = 21.1, P < .001$
Percentage male	65.8	40.0	
Age ± SD (y)	68.3 ± 7.5	74.2 ± 6.4	t = 7.0, $P < .001$
Min-Max	49-88	55-89	
Education ± SD (y)	12.8 ± 3.1	11.8 ± 2.9	t = 2.8, $P < .005$
Min-Max	8-20	7-20	
Mini-Mental State Examination[29] ± SD	29.0 ± 1.0 (follow-up, 28.9 ± 1.2)	24.1 ± 3.4 (follow-up, 22.6 ± 5.0)	t = 13.9, $P < .001$
Min-Max	25-30 (follow-up, 23-30)	16-30 (follow-up, 7-29)	

Three hundred seventy-four subjects fulfilled these criteria. Their characteristics are depicted in Table 1. Among the normal control (NC) subjects, 135 (36.1%) were apolipoprotein E ε4 carriers (ie, ε2/ε4, ε3/ε4, and ε4/ε4), and 233 (62.3%) were apolipoprotein E non-ε4 carriers (ie, ε2/ε2 or ε2/ε3, and ε3/ε3). Apolipoprotein E genotype information was missing because of technical problems in 6 subjects. No drug treatment or cognitive training was administered to the healthy controls between T_0 and T_1.

AD Patients With Mild Dementia

Data of 95 patients taken from the files of the Memory Clinic of the Geriatric University Hospital in Basel, Switzerland, with a diagnosis of probable AD according to the criteria outlined by the National Institute for Neurological and Communicating Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association[25] and DSM-IV criteria for AD[24] were included (Table 1). As part of their routine clinical workup, AD patients underwent extensive standardized neuropsychological testing and medical examinations, including neurological, neuroimaging, and laboratory investigations.[26] The average time interval between baseline and follow-up testing was 1.1 ± 0.24 years (range, 0.1-2.2 years). Patients were also screened for subjective symptoms of depression with the short form of the Geriatric Depression Scale (GDS).[27]According to this instrument, 79% of the patients had no depressive symptoms (GDS score, 0-4), and 16% suffered from mild (GDS score, 5-7) and 4% from moderate (GDS score, 8-10) depressive symptoms. None of the patients reported severe symptoms (more than 11 points).

The majority of patients had some kind of therapy between T_0 and T_1: 61 (64%) were treated with cholinesterase inhibitors. Eleven (12%) had cholinesterase inhibitor therapy plus weekly cognitive training,[28] and 3 (3%) had cognitive training only. Twenty patients (21%) had no therapeutic intervention. On average, patients visited the cognitive training sessions for 34.8 ± 27.15 weeks (range, 8-113 weeks) during their observation period. Within the AD sample, there was a group of patients with very mild cognitive decline (Mini-Mental State Examination [MMSE] ≥ 27)[29]; this subsample (n = 26; 13 females and 13 males; age at baseline, 73.9 ± 6.7; years of education, 13.0 ± 3.4; MMSE, 27.9 ± 0.8) was studied separately for some specific questions.

Assessment Tools

CERAD-NAB

NC subjects and AD patients were administered the validated German version of the CERAD-NAB[21] by experienced neuropsychologists or specially trained psychology students. This battery consists of subtests that cover the most commonly affected areas of cognitive functioning in AD patients. The tests are animal fluency, a modified Boston Naming Test (maximum score of 15), the MMSE (maximum score of 30), Word List-Learning (sum of three learning trials = Word List 1 + Word List 2 + Word List 3; maximum score of 30), Figures-Copy (maximum score of 11), Word List-Delayed Recall (maximum score of 10), Word List-Recognition (maximum of 100%), and Figures-Delayed Recall (maximum score of 11). Three new variables were created to obtain additional information: (1) Word List-Intrusions, which is the number of words mentioned over the 3 trials and the delayed recall that were not on the list; (2) Word List-Savings (%), which is the Word List-Delayed Recall/Word List-Trial 3 × 100;

and (3) Figures-Savings (%), which is the Figures-Delayed Recall/Figures-Copy × 100.

The CERAD-NAB can be usually completed in 20 to 45 minutes and is recommended for investigating different stages of dementia.[30] The English version of the test has been found to have good test-retest reliability, cross-center interrater reliability, and longitudinal validity.[30-32] For the German version, which is an exact translation of the English original, good to excellent discriminative validity (based on data from 1100 normal older subjects versus 150 AD patients) has been demonstrated.[33]

Statistical Analyses

Standardization of practice effects

To avoid the many difficulties inherent in reporting raw test scores of different tests, we decided to report our results as z scores throughout (ie, as a common, well-known metric for different kinds of raw scores established at different levels of ability.[9]

Practice effects were calculated by subtracting z scores at follow-up from z scores at baseline ($T_1 - T_0$) for each CERAD-NAB variable. The first step was therefore to establish z scores for a large sample of healthy aged subjects, which is described in detail elsewhere.[34] Briefly, raw scores of each CERAD-NAB variable were adjusted for age, gender, and years of education. Because linear regression models were to be used for this purpose, the most appropriate models were selected by applying Predicted Residual Sum of Squares[34] statistics. The Predicted Residual Sum of Squares statistic is the sum of squares of the predicted residuals for all observations where the predicted residual is the residual for an observation that results from dropping this observation from the model estimation. Each predicted residual is thus pretending to predict a "new" observation, and this idea is well suited for models that will be applied to future observations in practice. The most accurate model is the one with the smallest standard deviation of the predicted residuals. Because the residuals from several variables were not normally distributed (which is a prerequisite to establish z scores), the raw scores were first transformed to achieve a normal distribution of residuals. In the current analyses, we aimed at a normal distribution of residuals at least on the left-hand side because only low residuals are relevant for differentiating between healthy and diseased subjects. Once again, regression analyses were performed, and the regression model with the now smallest standard deviation of predicted residuals was selected. Variables were also tested for a possible age-related increase of variance. In a final step, z scores were calculated based on standardized residuals of the selected regression models (transformed) raw scores.

For practice effects, the procedure described previously here was applied to the z-score differences ($T_1 - T_0$) of each CERAD-NAB variable. In addition to demographic variables (age, education, and gender), baseline performance and all possible interactions between the explanatory variables were included in the regression analyses, providing 45 different regression models, among which the most appropriate one for each CERAD-NAB variable was selected.

Comparison between groups

The standardization procedure was performed within the NC group for 14 CERAD-NAB variables, and the results (formulae for z scores) were applied to the CERAD-NAB results of the AD patients. To facilitate comparison among measures, practice effects were also expressed in dimension-free units of effect size (ie, [z scores at follow-up − z score at baseline]/SD of z scores at baseline).

Stepwise binary logistic regression analyses with backward elimination (exclusion criterion, $P = .10$; inclusion criterion, $P = .05$) were performed on baseline scores and practice effects comparing NC subjects to AD patients and AD patients with an MMSE of 27 or more, respectively. A case was classified as AD if the predicted probability was greater than the proportion of true AD patients in the sample. This cutpoint was chosen to achieve similar levels of sensitivity and specificity.[35] For analyses between NCs and all AD patients, cutoffs were set at 0.2. For analyses between NCs and AD patients with MMSE ≥ 27, cutoffs were set at 0.07. The MMSE was discarded from the binary logistic analyses between NCs and AD patients with MMSE ≥ 27 because this variable had been used as a selection tool for AD patients.

Determination of cut-off scores for practice effects

Cutoff scores for practice effects were generated using receiver operating characteristic curves.[36] A receiver operating characteristic curve quantifies test accuracy and generates an empirically derived optimal cutoff

Table 2. Analyses of Changes Within the Normal Control Subjects (n = 374)

Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery Variable	Baseline (T_0) Mean ± SD	Follow-Up (T_1) Mean ± SD	Difference $T_1 - T_0$ ± SD	*P* Value (t Test)	Effect Size
Animal Fluency	0.06 ± 0.99	0.11 ± 0.99	0.05 ± 0.91	.23	0.05
Boston Naming Test	0.00 ± 0.95	0.04 ± 0.98	0.04 ± 1.00	.42	0.04
*MMSE	0.08 ± 0.95	0.04 ± 1.02	–0.04 ± 1.22	.57	–0.04
**Word List-Learning*	*–0.06 ± 0.99*	*0.35 ± 0.89*	*0.41 ± 0.99*	*< .001*	*0.41*
**Word List 1*	*0.05 ± 1.02*	*0.29 ± 0.95*	*0.24 ± 1.12*	*< .001*	*0.24*
**Word List 2*	*–0.10 ± 1.01*	*0.37 ± 0.90*	*0.47 ± 1.16*	*< .001*	*0.47*
**Word List 3*	*–0.06 ± 1.01*	*0.20 ± 1.02*	*0.26 ± 1.16*	*< .001*	*0.26*
Word List-Delayed Recall	*–0.04 ± 0.93*	*0.29 ± 0.90*	*0.33 ± 0.94*	*< .001*	*0.35*
Word List-Intrusions	*–0.03 ± 1.02*	*0.24 ± 0.84*	*0.27 ± 1.22*	*< .001*	*0.26*
**Word List-Savings*	*0.01 ± 1.06*	*0.20 ± 0.92*	*0.19 ± 1.29*	*.005*	*0.18*
Word List-Recognition	*0.02 ± 0.98*	*0.35 ± 0.81*	*0.33 ± 1.14*	*< .001*	*0.34*
Figures-Copy	*0.10 ± 0.94*	*–0.08 ± 0.97*	*–0.18 ± 1.24*	*.006*	*-0.19*
Figures-Delayed Recall	0.03 ± 0.90	0.11 ± 0.80	0.08 ± 0.98	.10	0.09
**Figures-Savings*	*0.00 ± 0.91*	*0.18 ± 0.93*	*0.18 ± 1.17*	*.003*	*0.20*

Note: Variables with an asterisk are those that remained in the final logistic regression (see text). *K* italics = significant change. Depicted are Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery z scores at baseline, follow-up, the differences, *P* values, and effect sizes (ie, $(T_1 - T_0)/SD_{T0}$).

score. It was decided a priori that sensitivity and specificity were equally important so that the optimal cut-off score was the point at which the sum of these indices was at a maximum.

Results

Study Sample

One NC subject had missing data on Animal Fluency at baseline, and in 2 AD patients, single baseline values were missing in the Boston Naming Test-15 items and Figures-Delayed Recall (and Figures-Savings). Two additional AD patients had single missing values at follow-up: Boston Naming Test-15 items and Figures-Delayed Recall. Because these missing values can be considered as missing at random, the data of these subjects were not discarded from the analyses.

Analyses Within the NC Sample

Standardization of the CERAD-NAB variable raw scores and practice effects

At T_0, all CERAD-NAB variables (raw scores) except Word List-Learning and Word Lists 1-3 needed to be transformed to be normally distributed. Age-related increase of variance was corrected for Word List-Savings and Figures-Delayed Recall. Practice effects (ie, z score at T_1 – z score at T_0) of the variables Word List-Recognition and Word List-Intrusions needed transformation to be normally distributed. Age-related increase of variance was corrected for Word List-Learning and Word List 3. With the stress criterion slightly relaxed, all scores could be modeled by either the main effects age, education, gender, and baseline performance or by adding some 2-way interactions between education, gender, and baseline performance to the main effects.

Comparison of cognitive test performance of NCs at T_0 and T_1

The performances at T_0 and T_1 as well as the test score differences (practice effects) and effect sizes between T_0 and T_1 are given in Table 2.

Applying Holm's sequentially rejective procedure[37] for 14 tests, significant improvements were seen on 9 of the 14 CERAD-NAB variables, and deterioration was seen in the variables MMSE and Figure-Copy (Table 2). According to effect sizes, practice effects were mild (ie, 0.2 – 0.47) in verbal learning (Word List-Learning, Word List-Intrusions, Word List-Delayed Recall, Word List-Recognition) and visual memory (Figures-Savings) variables. Animal Fluency, Boston Naming, and Figures-Delayed Recall did not show statistically significant practice effects.

Table 3. Analyses of Changes Within the Patients With Alzheimer's Disease (n = 95)

Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery Variable	Baseline (T_0) (mean ± SD)	Follow-Up (T_1) (mean ± SD)	Difference ($T_1 - T_0$ ± SD)	*P* Value (t-Test)	Effect Size	Practice Effects (Mean z Scores ± SD)
Animal Fluency	–1.50 ± 1.10	–1.62 ± 1.20	–0.12 ± 0.82	.16	–0.11	–1.08 ± 1.04
Boston Naming Test	–1.06 ± 1.39	–1.32 ± 1.50	–0.26 ± 1.02	.01	–0.19	–0.87 ± 1.27
Mini-Mental State Examination	*–2.81 ± 1.62*	*–3.40 ± 2.07*	*–0.59 ± 1.27*	*< .001*	*–0.36*	–2.69 ± 1.80
Word List-Learning	–2.49 ± 1.30	–2.61 ± 1.62	–0.13 ± 1.17	.30	–0.10	–2.36 ± 1.60
Word List 1	–1.69 ± 1.01	–1.76 ± 1.36	–0.06 ± 1.22	.63	–0.06	–1.60 ± 1.38
Word List 2	–2.06 ± 1.30	–2.08 ± 1.32	–0.02 ± 1.21	.86	–0.02	–2.25 ± 1.34
Word List 3	–2.28 ± 1.34	–2.51 ± 1.71	–0.24 ± 1.35	.09	–0.18	–1.72 ± 1.33
Word List-Delayed Recall	–2.28 ± 1.06	–2.28 ± 1.26	–0.00 ± 0.84	.99	–0.00	–1.96 ± 1.23
Word List-Intrusions	–0.53 ± 1.20	–0.49 ± 1.14	0.07 ± 1.42	.79	0.06	–0.32 ± 1.49
Word List-Savings	–1.61 ± 1.15	–1.59 ± 1.26	0.02 ± 1.32	.91	0.02	–1.72 ± 1.36
Word List-Recognition	–1.87 ± 1.32	–1.80 ± 1.53	0.04 ± 1.39	.64	0.03	–1.19 ± 1.69
Figures-Copy	–0.62 ± 1.49	–0.90 ± 1.72	–0.28 ± 1.33	< .05	–0.19	–0.68 ± 1.66
Figures-Delayed Recall	–1.52 ± 1.03	–1.68 ± 1.03	–0.15 ± 0.87	.10	–0.15	–1.84 ± 1.16
Figures-Savings	–1.73 ± 1.18	–1.91 ± 1.24	–0.17 ± 1.10	.14	–0.14	–1.93 ± 1.21

Note: *K* italics = significant change. Depicted are Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery z scores at baseline, follow-up, the differences, *P* values, and effect sizes (ie, $(T_1 - T_0)/SD_{T0}$), and practice effects.

Analysis Within the AD Sample

Comparison of cognitive test performance of AD patients at T_0 and T_1

Independent t tests revealed no significant differences and no trends ($P > .1$) of baseline or practice effect on performance between drug-treated (n = 72) and untreated (n = 23) AD patients; therefore, drug-treated and untreated AD patients were analyzed together. After applying Holm's sequentially rejective procedure, a significant decline from T_0 to T_1 among all AD patients—and also for the subsample of AD patients with MMSE scores of 27 or more (results not shown)—was found on the MMSE, with effect sizes of –0.36 and –0.99, respectively (Table 3). There was some decline in performance at T_1 (data not significant) among the AD patients in all variables except for Word-List Intrusions, Word-List Savings, and Word-List Recognition. For AD patients with MMSE scores of 27 or more, a decline was only found on the variables Boston Naming Test, Figures Copy, and Figures Savings. A slight improvement at T_1 (not statistically significant) was found for the remaining variables.

We also studied whether the length of the time interval between T_0 and T_1 (which varied across the AD patients) was statistically related to changes of cognitive performance in the patient group, but we found no significant association (data not shown).

Comparisons Between the NC and the AD Samples

Stepwise binary logistic regression analyses for baseline scores and practice effects

Binary logistic regression analyses using baseline scores between NC subjects and AD patients revealed that the CERAD-NAB subtests Animal Fluency, MMSE, Word List-Learning, Word Lists 1-3, and Word List-Delayed Recall distinguished NCs from AD patients most efficiently. The specificity rate was 94.9%, and the sensitivity rate was 95.7%, with an overall correct classification rate of 95.1%.

Binary logistic regression analyses based on practice effects (T_1 minus T_0) revealed that the CERAD-NAB subtests MMSE, Word List-Learning, Word Lists 1-3, Word List-Savings, and Figures-Savings distinguished NCs from AD patients most efficiently; 93.3% of the NCs and 86.8% of the AD patients were correctly classified, yielding an overall correct classification rate of 92.0%. The rate of the explained variance according to Nagelkerke's estimate for R^2 was 80.9%.

Among the NCs, 331 were correctly classified by baseline scores and practice effects. Twenty two were correctly classified by baseline scores but were wrongly classified through practice effects, and 3 subjects were wrongly classified by both approaches. Among the AD patients, 81 were correctly classified by

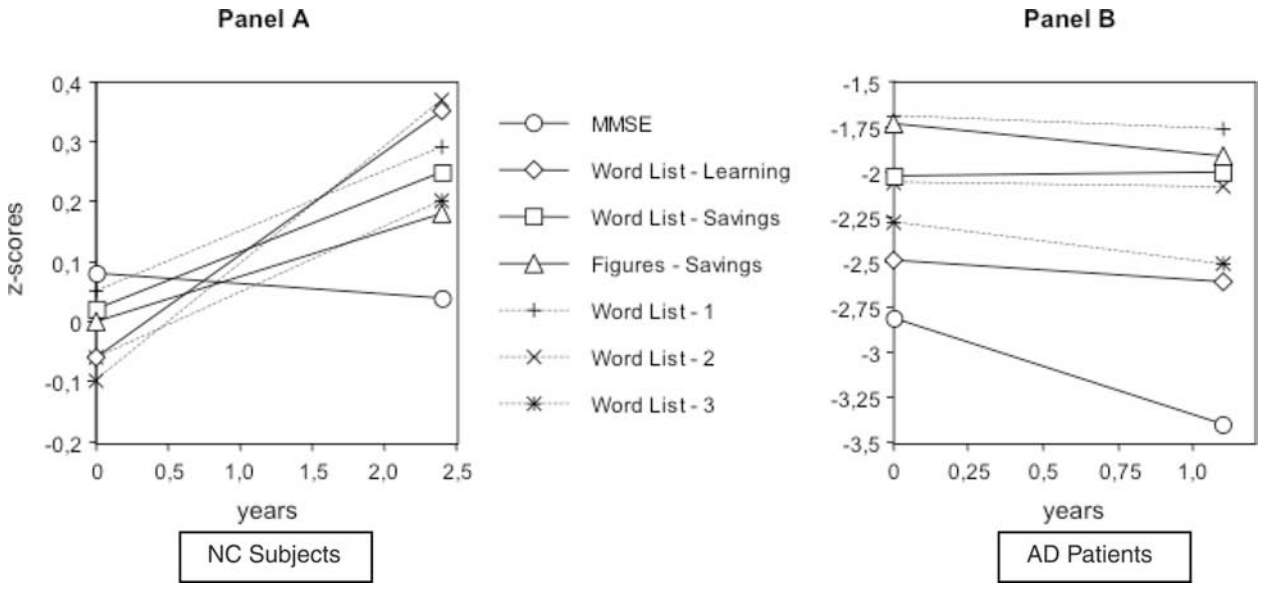

Figure 1. Z scores of AD patients (A) and normal control subjects (B) at 2 consecutive assessments. Presented are only those variables that remained in the final logistic regression to separate normal control subjects and patients with Alzheimer's disease based on their change scores.

baseline scores and practice effects. Nine were correctly classified by baseline scores but wrongly through practice effects, and 4 subjects were wrongly classified by both approaches. None of the AD patients was wrongly classified by baseline scores but was correctly allocated by practice effects. According to the McNemar test, this represents a significant result ($P = .004$) in favor of baseline scores.

Figure 1 illustrates NC subjects' (Figure 1A) and AD patients' (Figure 1B) CERAD-NAB baseline and follow-up z scores for those variables that remained in the final logistic regression equation.

For AD patients with very mild dementia (MMSE scores of 27 or more), the specificity rate was 88.2%. The sensitivity rate 88.5%, with an overall classification rate of 88.2% using baseline scores. Animal Fluency, Word List-Learning, Word Lists 1-3, Word List-Delayed Recall, and Figures-Delayed Recall distinguished NCs from AD patients most efficiently. Binary logistic regression analyses based on practice effects (T_1 minus T_0) revealed that the CERAD-NAB subtests Word List-Learning, Word Lists 1-3, Word List-Savings, and Figures-Savings distinguished NCs from AD patients most efficiently; 84.9% of the NCs and 80.8% of the AD patients were correctly classified, yielding an overall correct classification rate of 84.7%.

Among the healthy control subjects, 289 were correctly classified by baseline scores and practice effects. Thirty nine were correctly classified by baseline scores but wrongly through practice effects, and 17 subjects were wrongly classified by both approaches. Among the AD patients 20 were correctly classified by baseline scores and practice effects, 3 were correctly classified by baseline scores but wrongly through practice effects, and 2 subjects were wrongly classified by both approaches. McNemar tests revealed no significant differences between the 2 approaches (baseline scores versus practice effects).

Determining optimal cutoff scores for practice effects

Based on the a priori taken decision that sensitivity and specificity were equally important, receiver operating characteristic curves were constructed to determine the optimal cutoff z scores for the different CERAD-NAB subtests discriminating AD patients from NCs. Table 4 displays the findings in terms of optimal cutoff scores, sensitivity, specificity, and correct classification rates; the latter varied between 50% (Figures-Copy) and 86% (MMSE).

Table 4. Empirical Thresholds for the Discrimination Between Normal Control Subjects and Alzheimer's Disease Patients Through "Practice Effects"

Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery Variables	Cutoff Score (z Score)	Sensitivity (%)	Specificity (%)	Correct Classification Rate (%)
Animal Fluency	–0.50	67.5	78.0	75.8
Boston Naming Test	–0.17	60.5	74.7	71.8
Mini-Mental State Examination	–1.18	88.7	85.7	86.3
Word List-Learning	–1.25	91.7	76.9	79.9
Word List 1	–0.58	72.6	79.1	77.8
Word List 2	–0.96	82.3	82.4	82.4
Word List 3	–1.28	92.5	67.0	72.2
Word List-Delayed Recall	–1.15	87.9	75.8	78.2
Word List-Recognition	–0.51	83.9	74.7	76.6
Word List-Savings	–1.30	92.2	69.2	73.9
Word List-Intrusions	–0.45	78.5	52.7	57.9
Figures-Copy	–1.18	88.7	40.7	50.4
Figures-Delayed Recall	–1.53	94.9	73.6	77.9
Figures-Savings	–1.51	94.9	74.7	78.8

Comparison of practice effects between NCs and AD patients

Table 5 shows mean z-score differences and corresponding raw scores between NCs and AD patients. Compared with the entire sample of AD patients, NCs showed significantly superior practice effects on all 14 variables (applying Holm's sequentially rejective procedure). Differences were most evident on the variables MMSE, Word List-Learning (especially Word List 2), and Word List-Delayed Recall. For the subgroup of AD patients with MMSE scores of 27 or more, practice effects were also significantly less than seen in NCs on most variables except for the Boston Naming Test, Word List-Recognition, Word List-Savings, Word List-Intrusions, and Figures Copy.

Discussion

After an average test-retest interval of 2.4 years, healthy aged subjects (n = 374; mean age, 68 years) demonstrated significant gains of performance on 9 of 14 parameters assessed by means of a widely used neuropsychological test battery, the CERAD-NAB. Improvements from baseline to follow-up were most evident on variables derived from a verbal learning task (Word List-Learning, Word List 2, Word List-Delayed Recall, and Word List-Recognition), whereas performance gains were less pronounced or even absent for variables with so-called ceiling effects, that is, the Boston Naming Test, Figures Copy (significant deterioration), and MMSE, which hardly offered room for improvement: 91% of the healthy control subjects reached MMSE scores of 28 or more at baseline, 89% at least 10 of maximal 11 points on Figures Copy, and 92% at least 13 of maximal 15 points on the Boston Naming Test at baseline. Because the test results were adjusted for possible confounding variables such as age, gender, education, and baseline values, and as it is unlikely that specific test items were memorized over 2.4 years, the improvements from T_0 to T_1 are most likely to be accounted for by what is commonly referred to as "test sophistication."[17] This term comprises a number of factors that eventually help subjects to improve their performance on repeated tests, for example, knowledge about test-taking procedures, effective test-taking strategies, a reduced sense of novelty, and fewer feelings of fear or nervousness at retest. Diminished or missing practice effects could then be indicative of an as yet unnoticed or denied decline of cognitive adaptability or mental flexibility.

In contrast to the group of NCs who are volunteers in a scientific study, the patients with mostly mild AD attended the Basel Memory Clinic mainly for diagnostic purposes and consultation; different from the NCs, they did not improve on a majority of the neuropsychological performance parameters when retested after an average interval of about one year. Repeated neuropsychological testing is not a routine

Table 5. Comparison of Mean Differences of Practice Effects (z Scores and Corresponding Raw Scores) Between Normal Control Subjects and Alzheimer's Disease Patients and Alzheimer's Disease Patients With an Mini-Mental State Examination Score of 27

Consortium to Establish a Registry for Alzheimer's Disease-Neuropsychological Assessment Battery Variables	Normal Controls Versus Alzheimer's Disease				Normal Controls Versus Alzheimer's Disease With Mini-Mental State Examination ≥ 27			
	Mean Difference	Corresponding Raw Scores	95% Confidence Interval	*P* Value	Mean Difference	Corresponding Raw Scores	95% Confidence Interval	*P* Value
Animal Fluency	–1.08	–4.4	–1.31, –0.85	< .001	–0.61	–2.5	–1.01, –0.21	.003
Boston Naming Test	–0.87	–0.9	–1.15, –0.59	< .001	–0.39	–0.4	–0.79, 0.01	.06
Mini-Mental State Examination	–2.69	–3.7	–3.07, –2.31	< .001	–1.35	–1.8	–1.75, –0.95	< .001
Word List-Learning	–2.36	–5.9	–2.70, –2.02	< .001	–1.11	–2.8	–1.66, –0.55	< .001
Word List 1	–1.59	–1.9	–1.89, –1.29	< .001	–0.62	–0.7	–1.02, –0.21	.003
Word List 2	–2.25	–2.6	–2.54, –1.96	< .001	–1.43	–1.6	–1.82, –1.03	< .001
Word List 3	–1.72	–2.1	–2.01, –1.44	< .001	–0.86	–1.1	–1.24, –0.48	< .001
Word List-Delayed Recall	–1.97	–3.2	–2.24, –1.70	< .001	–1.01	–1.6	–1.41, –0.60	< .001
Word List-Recognition	–1.19	*	–1.55, –0.83	< .001	–0.44	*	–0.98, 0.10	.11
Word List-Savings	–1.71	–37.6	–2.01, –1.42	< .001	–0.79	–17.3	–1.41, –0.17	.015
Word List-Intrusions	–0.35	0.2	–0.67, –0.03	.033	–0.98	–0.1	–0.70, 0.50	.11
Figures-Copy	–0.68	–0.7	–1.03, –0.33]	< .001	0.05	0.0	–0.36, 0.45	.83
Figures-Delayed Recall	–1.84	–4.4	–2.08, –1.61	< .001	–0.96	–2.3	–1.53, –0.40	< .001
Figures-Savings	–1.93	*	–2.16, –1.69	< .001	–1.11	*	–1.65, –0.57	< .001

*Calculation of raw changes involves derivatives of transformations. For two variables, Word List-Recognition and Figures-Savings, transformations are not differentiable because of piecewise fit. Change in raw data could therefore not be calculated for these variables.

procedure at the Basel Memory Clinic; the patients included in this sample were followed up either (1) to confirm a previous diagnosis or (2) to document any potential benefit from treatment or (3) for other specific individual reasons. These AD patients with generally mild dementia, and in particular, the subgroup with MMSE scores of 27 or more showed either no significant improvement or some decline on a number of performance parameters, notably the MMSE. A direct comparison of change scores from T_0 to T_1 between the two groups of subjects revealed significant differences in favor of the NCs with regard to the majority of CERAD-NAB parameters.

Given the marked differences between NCs and AD patients regarding their abilities to enhance performance from one test occasion to the next, it appeared logical to investigate whether the observed differences in practice effects had any diagnostic potential, that is, whether differences in performance on the CERAD-NAB from T_0 to T_1 could be reliably used to differentiate between NCs and patients with AD. Stepwise logistic regression analysis using the differences in performance between T_0 and T_1 identified a number of CERAD-NAB measures, based on which 92.0% of all subjects (93.3% of NCs and 86.8% of AD patients) were correctly classified. Although this represents an excellent result, it was inferior to the discrimination obtained between the NCs and AD patients when the baseline CERAD-NAB findings were used (95.1% correct classifications; 94.9% of NCs, 95.7% of AD patients correctly classified). Nevertheless, the analyses of practice effects are definitely interesting in that the group of older NCs as a whole showed a significantly different development of their cognitive performance over time when compared with patients suffering from mild AD. As predicted, the NCs displayed improved performance on all neuropsychological measures at retesting after 2.4 years (9 of 14 criteria statistically significant), whereas the AD patients showed some, although not consistently, significant decline on all parameters after a much shorter retest interval (1.1 years). The fact that the decline of performance seen in the AD patients was generally moderate should be attributed (1) to the short test-retest interval and (2) to the fact that the AD patients were in mild or even very mild stages of their disease. The differences in change of cognitive performance between the 2 groups were

statistically significant on 13 out of 14 neuropsychological measures studied.

Within the AD group, a comparison between the two approaches (baseline scores versus practice effects) favored the use of baseline scores. For AD patients with very mild dementia (MMSE ≥ 27), both baseline scores and practice effects were similarly efficient. Additional analyses of the NC subjects and AD patients incorrectly classified through baseline scores or practice effects, respectively, revealed no specific features that could explain why they had been wrongly allocated (data not shown). It might be that drug treatment received by these patients between baseline and follow-up contributed to some of the wrong classifications.

In summary, the analyses presented indicate that (1) normal aged controls showed improvements, to a great extent statistically significant, on the variables of the CERAD-NAB after a retest interval of more than 2 years; (2) AD patients with generally mild (and even very mild) dementia failed to show the improvement in test performance observed in NCs, despite a much shorter retest interval; (3) discrimination between groups of NCs and AD patients was excellent for both CERAD-NAB baseline values and change (T_1 minus T_0) scores; (4) however, change scores were not diagnostically superior to baseline performance values.

Taken together, these findings suggest that repeated testing with the CERAD-NAB does not generally add to improved diagnostic accuracy for mild and very mild AD and cannot, therefore, be recommended as a routine clinical procedure.

Acknowledgment

This study was supported by the Swiss National Science Foundation (grant 3200-049107.96) and by an unconditioned scientific grant from the Novartis Foundation. We gratefully acknowledge the help and support of all patients and volunteers as well as the staff of the Memory Clinic, Basel, Switzerland.

References

1. Ott A, Breteler MM, van Harskamp F, et al. Prevalence of Alzheimer's disease and vascular dementia: association with education: the Rotterdam study. *BMJ*. 1995; 310:970-973.
2. Katzman R, Kawas C. The epidemiology of dementia and Alzheimer disease. In Terry RD, Katzman R, Bick KL, eds. *Alzheimer Disease*. New York, NY: Raven Press; 1994:105-122.
3. Bäckman L, Jones S, Berger A-K, Laukka EJ, Small BJ. Multiple cognitive deficits during transition to Alzheimer's disease. *J Intern Med*. 2004:256:195-204.
4. Small BJ, Mobly JL, Jonsson Laukka E, Jones S, Bäckman L. Cognitive deficits in preclinical Alzheimer's disease. *Acta Neurol Scand*. 2003;107:29-33.
5. Howieson DB, Dame A, Camicioli R, Sexton G, Payami H, Kaye JA. Cognitive markers preceding Alzheimer's dementia in the healthy oldest old. *J Am Geriatr Soc*. 1997;45:584-589.
6. Chen P, Ratcliff G, Belle SH, et al. Patterns of decline in pre-symptomatic Alzheimer's Disease: a prospective community study. *Arch Genet Psychiatry*. 2001;58:853-858.
7. Collie A, Maruff P, Shafiq-Antonacci R, et al. Memory decline in healthy older people. *Neurology*. 2001;56: 1533-1538.
8. Winblad B, Palmer K, Kivipelto M, et al. Mild cognitive impairment: beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *J Intern Med*. 2004;256:240-246.
9. Lezak MD, Howieson D-B, Loring D-W, Hannay H-J, Fischer J-S. *Neuropsychological Assessment*. 4th ed. New York, NY: Oxford University Press; 2004:Xiv, 1016.
10. Longstreth LE, Alcorn MB. Susceptibility of Wechsler Spatial Ability to experience with related games. *Educ Psychol Meas*. 1990;50:1-6.
11. McCaffrey RJ, Westervelt HJ. Issues associated with repeated neuropsychological assessments. *Neuropsychol Rev*. 1995;5:203-221.
12. Rabbitt P, Diggle P, Smith D, Holland F, McInnes L. Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia*. 2001;39: 532-543.
13. Frank R, Wiederholt WC, Kritz-Silverstein D, Salmon DP, Barrett-Connor E. Effects of sequential neuropsychological testing of an elderly community-based sample. *Neuroepidemiology*. 1996;15:257-268.
14. Lowe C, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. *Neuropsychologia*. 1998;36:915-923.
15. Mitrushina M, Satz P. Effect of repeated administration of a neuropsychological battery in the elderly. *J Clin Psychol*. 1991;47:790-801.
16. Basso MR, Bornstein RA, Lang JM. Practice effects on commonly used measures of executive function across twelve months. *Clin Neuropsychol*. 1999;13:283-292.
17. Anastasi A. Coaching, test sophistication, and developed abilities. *Am Psychol*. 1981;36:1086-1093.
18. Temkin NR, Heaton RK, Grant I, Dikmen SS. Detecting significant change in neuropsychological test performance: a comparison of four models. *J Int Neuropsychol Soc*. 1999;5:357-369.

19. Monsch AU, Thalmann B, Schneitter M, et al. The Basel study on the elderly's search for preclinical cognitive markers of Alzheimer's disease. *Neurobiol Aging.* 2000;21:31.
20. Widmer LK, Stähelin HB, Nissen C, da Silva A. *Venen-, Arterien-Krankheiten, koronare Herzkrankheit bei Berufstätigen: Prospektiv-epidemiologische Untersuchung. Basler Studie I-III 1959-1978.* Bern, Switzerland: Hans Huber; 1981.
21. Thalmann B, Monsch AU, Schneitter M, et al. The CERAD neuropsychological assessment battery (CERAD-NAB): a minimal dataset as a common tool for German-speaking Europe. *Neurobiol Aging.* 2000;21:30.
22. Thalmann B, Spiegel R, Stähelin HB, et al. Dementia screening in general practice: simplified scoring for the Clock Drawing Test. *Brain Aging.* 2002;2:36-43.
23. Kühner C. *Fragebogen zur Depressionsdiagnostik Nach DSM-IV (FDD-DSM-IV).* Göttingen, Germany: Hogrefe; 1997.
24. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* 4th ed. Washington DC: American Psychiatric Association; 1994.
25. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. 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.* 1984;34:939-944.
26. Monsch AU, Foldi NS, Ermini-Fünfschilling DE, et al. Improving the diagnostic accuracy of the Mini-Mental State Examination. *Acta Neurol Scand.* 1995;92:145-150.
27. Sheikh JI, Yesavage JA. Geriatric *Depression Scale (GDS): Recent Evidence and Development of a Shorter Version: Clinical Gerontology: A Guide to Assessment and Intervention, NY: The Haworth Press;* 1986:165-173.
28. Ermini-Fünfschilling D, Meier D. Gedächtnistraining: Wichtiger Bestandteil der Milieutherapie bei seniler Demenz. *Zeitschrift Gerontol Geriatrie.* 1995;28:190-194.
29. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12:189-198.
30. Morris JC, Heyman A, Mohs RC, et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD): I: clinical and neuropsychological assessment of Alzheimer's disease. *Neurology.* 1989;39:1159-1165.
31. Welsh KA, Butters N, Hughes J, Mohs R, Heyman A. Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Arch Neurol.* 1991;48:278-281.
32. Welsh KA, Butters N, Hughes J, Mohs RC, Heyman A. Detection and staging of dementia in Alzheimer's disease. *Arch Neurol.* 1992;49:448-452.
33. Aebi C. Validierung der neuropsychologischen Testbatterie CERAD-NP: Eine Multi-Center Studie [PhD thesis]. Basel, Switzerland: University of Basel; 2002.
34. Berres M, Monsch AU, Bernasconi F, Thalmann B, Stähelin HB. Normal ranges of neuropsychological tests for the diagnosis of Alzheimer's disease. *Stud Health Technol Inform.* 2000;77:195-199.
35. Hosmer DW, Lemeshow S. *Binary Logistic Regression.* 2nd ed. New York, NY: Wiley; 2000.
36. Armitage P, Berry G. *Statistical Methods in Medical Research.* 3rd ed. Boston, Mass: Blackwell Scientific; 1994.
37. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6:65-70.