Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10 (2018) 372-381

Diagnostic Assessment & Prognosis

Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects

Jeremy A. Elman[a,b,*], Amy J. Jak[a,c], Matthew S. Panizzon[a,b], Xin M. Tu[d], Tian Chen[e], Chandra A. Reynolds[f], Daniel E. Gustavson[a,b], Carol E. Franz[a,b], Sean N. Hatton[a,b], Kristen C. Jacobson[g], Rosemary Toomey[h], Ruth McKenzie[h], Hong Xian[i,j], Michael J. Lyons[h], William S. Kremen[a,b]

[a]*Department of Psychiatry, University of California San Diego, La Jolla, CA, USA*
[b]*Center for Behavior Genetics of Aging, University of California San Diego, La Jolla, CA, USA*
[c]*Psychology Service, VA San Diego Healthcare System, La Jolla, CA, USA*
[d]*Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA*
[e]*Department of Mathematics and Statistics, University of Toledo, Toledo, OH, USA*
[f]*Department of Psychology, University of California Riverside, Riverside, CA, USA*
[g]*Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, IL, USA*
[h]*Department of Psychology, Boston University, Boston, MA, USA*
[i]*Department of Statistics, St Louis University, St Louis, MO, USA*
[j]*Research Service, VA St Louis Healthcare System, St Louis, MO, USA*

Abstract

Introduction: Longitudinal testing is necessary to accurately measure cognitive change. However, repeated testing is susceptible to practice effects, which may obscure true cognitive decline and delay detection of mild cognitive impairment (MCI).

Methods: We retested 995 late-middle-aged men in a ~6-year follow-up of the Vietnam Era Twin Study of Aging. In addition, 170 age-matched replacements were tested for the first time at study wave 2. Group differences were used to calculate practice effects after controlling for attrition effects. MCI diagnoses were generated from practice-adjusted scores.

Results: There were significant practice effects on most cognitive domains. Conversion to MCI doubled after correcting for practice effects, from 4.5% to 9%. Importantly, practice effects were present although there were declines in uncorrected scores.

Discussion: Accounting for practice effects is critical to early detection of MCI. Declines, when lower than expected, can still indicate practice effects. Replacement participants are needed for accurately assessing disease progression.

Keywords: Practice effects; Repeat testing; Serial testing; Longitudinal testing; Mild cognitive impairment; Cognitive change

1. Introduction

Longitudinal assessments are necessary for directly measuring cognitive change over time to track disease progression from cognitively normal to mild cognitive impairment (MCI) or MCI to Alzheimer's disease (AD), and for assessing efficacy of therapeutic interventions [1,2]. Because the pathological process begins decades before the onset of AD, it is widely agreed that early identification is of enormous importance [3]. However, repeat testing is susceptible to practice effects [4,5], and failure to account for practice effects may obscure cognitive declines and delay detection of conversion to MCI or AD [6,7].

The authors J.A.E. and A.J.J. contributed as joint first authors.
The authors have declared that no conflict of interest exists.
*Corresponding author. Tel.: (858) 534-6842; Fax: (858) 822-5856.
E-mail address: jaelman@ucsd.edu

https://doi.org/10.1016/j.dadm.2018.04.003

Here, we assessed practice effects on neuropsychological testing and their impact on diagnosis of MCI in late-middle-aged adults. Practice effects are typically defined as improvements in performance due to prior exposure to a test as opposed to frank cognitive improvements [5,8]. However, in longitudinal studies, the absence of practice effects over time could signal cognitive decline as opposed to cognitive stability [9]. In midlife and later life, when normative declines are expected, stable performance, and even declines, could still reflect the contribution of practice effects. This situation is particularly problematic for studies that use age-based norms to diagnose individuals with cognitive impairment. Consider, for example, two individuals with similar characteristics who have identical cognitive test scores just above threshold for an MCI diagnosis, the only difference is that one individual is being tested for the first time, whereas the other has taken these tests before. We can infer that the second individual may actually have more impairment, but the effects of practice are artificially increasing their test scores, keeping them above threshold. This individual may have been diagnosed as having MCI had they been taking the test for the first time and not received the benefits of practice. Clinically, this scenario is becoming more relevant as AD drug intervention trials shift toward secondary prevention strategies that rely on early identification. Failure to correct for practice effects may result in underdiagnosis or delays in detecting MCI.

Shorter test-retest intervals are significantly associated with an increased magnitude of practice effects [8,10,11]. However, practice effects have been found across intervals of over 5 years, and it has been estimated that it may take at least 7 years for practice effects to decrease to zero in adults aged 18–58 years [12,13]. Alternate test forms do not solve the problem because they do not fully remove practice effects [14,15] and they introduce test differences as yet another factor that affects performance. Selective attrition is an additional concern. Because returnees usually represent a healthier or higher performing subgroup, they would be expected to score higher at follow-up than the overall sample at baseline [1,16]. In case-control studies, the control group may be used to gauge practice effects, but that only allows for assessment of relative change in cases versus controls. If, however, there is no explicit control group and the goal is to determine when someone meets criteria for a diagnosis of MCI, cutoff scores for cognitive impairment at the point of diagnosis must be made. If there are practice effects, then those cutoff scores should be modified. The standard approaches for gauging practice effects do not provide a way to adjust the impairment threshold.

In contrast, inclusion of replacement participants—individuals who complete their baseline testing visit at the same time and age as the initial sample's follow-up visit—provides an optimal strategy to calculate practice and attrition effects [13]. Rönnlund et al. [13] used this approach to examine practice effects on tests of episodic and semantic memory. Here, we applied this approach to investigate practice effects across multiple cognitive domains over a 6-year interval during late midlife. We hypothesized that adjusting for practice effects would result in an increased rate of MCI at follow-up, suggesting that MCI cases are being missed when practice effects are not taken into account. In addition, because we previously found that different episodic memory tests showed different patterns of change across time [17], we examined whether practice effects contributed to these differences.

2. Methods

2.1. Participants

Participants were from waves 1 and 2 of the Vietnam Era Twin Study of Aging (VETSA) [18]. VETSA participants comprise a national, community-dwelling sample of male-male twins who are similar to American men in their age range with respect to health and lifestyle characteristics based on Center for Disease Control and Prevention data [19]. All served in the military sometime between 1965 and 1975, but nearly 80% reported no combat exposure. Detailed descriptions of the sample composition and method of ascertainment have been reported elsewhere [20,21].

The current analysis included 1220 individuals tested at wave 1 (mean age = 55.88 years, standard deviation [SD] = 2.5). Of these, 225 were dropouts and 995 (82%) returned for wave 2 (returnees; mean age = 61.54 years, SD = 2.4). At wave 2, 170 attrition replacements (ARs) of a similar age as the returnees were tested for the first time (mean age = 62.67 years, SD = 2.3). ARs were recruited from the same twin registry as the other participants.

The study was approved by the institutional review boards at the participating institutions.

2.2. Demographics

The estimation of practice effects based on group means of returnees and ARs assumes the samples are well-matched on measures that may cause systematic differences in test performance. Although participants were recruited from a random sample, it is still possible that group differences exist. We found some small but significant differences between groups (see Table 1). The returnee group had a higher percentile score on the general cognitive ability (GCA) test that was administered at an average age of 20 years (61.4 vs. 54.3; $t_{(225.79)} = 3.765$, $P < .001$). These scores are approximately equivalent to intelligence test (IQ) scores of 104.4 versus 101.6. Section 2.3 contains a description of the GCA test. The returnee group also had a higher average education (13.9 years vs. 13.4 years, $t_{(222.54)} = 3.765$, $P < .001$). The AR group was older than returnees (62.67 vs. 61.54; $t_{(238.98)} = 5.880$, $P < .001$). The percentage of apolipoprotein E ε4 (*APOE*-ε4) carriers was similar between groups (29.4% vs. 25.8%; $\chi^2(1) = 2.16$, $P = .142$).

Table 1
Demographic characteristics of the sample

Measure	Dropouts	Returnees	Attrition replacements
N	225	995	170
Age, in years	Wave 1: 56.23 (2.5)*	Wave 1: 55.80 (2.5) Wave 2: 61.54 (2.4)	Wave 2: 62.67 (2.3)**
Education	13.70 (2.1)	13.87 (2.09)	13.44 (2.2)*
Age 20 GCA	59.95 (23.11)	61.40 (22.1)	54.27 (22.9)**
APOE-ε4+	58 (25.8%)	293 (29.4%)	37 (21.8%)

NOTE. GCA = general cognitive ability based on the Armed Forces Qualification Test (AFQT). Standard deviations (SDs) are reported in parentheses (except when reporting percentages).

NOTE. *$P < .05$ and **$P < .001$ significance of comparison with returnees.

2.3. Cognitive measures

A set of 23 measures grouped into seven specific cognitive abilities was assessed for practice effects. Scores at both waves were converted to z-scores based on the means and SDs of all participants' wave 1 scores. When there was more than one measure of a cognitive ability, we generated composites by taking the average of individual z-score measures. Signs of some tests (e.g., reaction time) were flipped so that higher scores always indicated better performance. The tests and scores comprising the seven specific cognitive abilities are shown in Table 2.

In addition to the specific abilities, general ("premorbid") cognitive ability (GCA) in young adulthood (average age of 20 years) was measured with the Armed Forces Qualification Test (AFQT), a 100-item multiple-choice paper-and-pencil test administered just before military induction [22]. The AFQT is highly correlated (r = .84) with standard IQ measures [23]. Percentile scores were transformed to a normal distribution using the probit function.

2.4. Adjustment for GCA and education

Owing to the difference in age 20 GCA between returnees and ARs, practice effects might be inflated. The higher level of performance in the returnees at follow-up could be partly due to their higher average GCA. To account for this difference, we regressed out the transformed age 20 AFQT score from all test scores at waves 1 and 2. The intercept was added back into the residuals to maintain mean-level information in each test score.

Our primary analyses focused on the scores adjusted for age 20 GCA because this is a direct measure of early adult cognitive ability for which education is often used as a proxy. Elsewhere, we have shown a full distribution of age 20 AFQT scores even among participants who all had exactly 12 years of education [24]. Therefore, adjusting by age 20 AFQT scores provides more accurate matching than adjusting by education. However, because direct measures of early-life GCA are rarely available, we also created a data set adjusted for education using the same methods as described previously. The practice effects generated from this data set may provide a useful reference for future studies.

2.5. Estimation of practice effects

Practice effects were calculated according to the methods of Rönnlund et al. [13]. If there are no attrition effects, this method assumes that if two groups are drawn from the same population and the only difference is that one group has taken a test before and the other is being tested for the first time, any group difference can be attributed to the effect of practice. Therefore, we compare the mean of wave 2 scores between returnees taking the test for a second time and the mean score of age-matched ARs taking the test for the first time. In reality, this difference reflects a combination of practice effects and selective attrition effects. That is, the individuals who returned for follow-up may represent a better performing or healthier subset of the overall baseline sample. This approach therefore calculates the attrition effect as the difference in the mean score of returnees at wave 1 and the mean score of all individuals at wave 1. The practice effect of each test was then calculated as the difference score minus the attrition effect. In summary,

$$\text{Difference score} = \text{Returnees}_{T2} - \text{Replacements}_{T2}$$
$$\text{Attrition effect} = \text{Returnees}_{T1} - \text{All}_{T1}$$
$$\text{Practice effect} = \text{Difference score} - \text{Attrition effect}$$

Where $Returnees_{T2}$ is the mean score of returnees at wave 2, $Replacements_{T2}$ is the mean score of ARs at wave 2, $Returnees_{T1}$ is the mean score of returnees at wave 1, and All_{T1} is the mean score of the entire group at wave 1. The resulting value of the practice effect may be interpreted as the number of SD unit change in wave 1 scores that is expected simply due to repeat testing. A positive value represents an expected *increase* in performance resulting from practice. These values were then subtracted from returnees' wave 2 scores to obtain corrected test scores, that is, the score that would be expected if the individual were taking the test for the first time. Importantly, this approach simply removes the mean expected change (or "group-level practice effects"), but it does not remove any individual differences in change (or "individual practice effects"). Therefore, individual variability can still be assessed with any other methods.

Standard errors were estimated using bootstrap resampling (10,000 resamples). In each resampling, a random sample with replacement was drawn from each group with a sample size matching the original group.

Table 2
Means, standard deviation (SD), and estimates practice effects

Measure	Dropouts time 1 Mean (SD)	Returnees time 1 Mean (SD)	Returnees time 2 Mean (SD)	Attrition replacements time 2 Mean (SD)	Practice effect	SE	*P*
Visual-spatial					0.318	0.063	<.001
Mental Rotation: Total Correct Part 1 [44]	0.06 (0.99)	−0.02 (0.99)	−0.01 (1.00)	−0.27 (0.91)	0.275	0.077	<.001
Hidden Figures Total Correct All Parts [45]	−0.02 (1.03)	0.00 (0.99)	−0.20 (0.93)	−0.57 (0.87)	0.360	0.074	<.001
Executive					0.085	0.059	.1404
D-KEFS Trails 4 time (adjusted for Trails 2 and 3, log transformed)	−0.09 (0.95)	–0.01 (1.01)	−0.07 (0.98)	−0.17 (1.14)	0.083	0.093	.339
D-KEFS Category Switching Accuracy (adjusted for verbal fluency)	−0.02 (1.04)	−0.01 (0.99)	−0.04 (1.02)	−0.16 (1.18)	0.106	0.096	.222
Stroop: Interference Score (adjusted for color and word) [46]	−0.14 (0.97)	0.03 (1.00)	−0.26 (1.05)	−0.358 (1.08)	0.064	0.090	.4759
Short-term/working memory					0.226	0.046	<.001
WMS-III Digit Span: Forward Raw Score	−0.01 (1.04)	0.01 (0.99)	−0.05 (0.98)	−0.21 (1.03)	0.152	0.084	.071
WMS-III Digit Span: Backward Raw Score	0.01 (1.04)	0.00 (0.99)	−0.02 (0.96)	−0.28 (0.77)	0.258	0.067	.001
WMS-III Spatial Span: Total Trials Passed	−0.05 (1.05)	0.03 (1.00)	−0.11 (1.01)	−0.51 (0.96)	0.388	0.080	<.001
WMS-III Spatial Span: Total Trials Passed Backward	−0.12 (0.99)	0.03 (1.00)	−0.19 (0.99)	−0.49 (0.91)	0.268	0.077	<.001
WMS-III Letter-Number Sequencing: Total Score for Trials Passed	−0.04 (0.98)	0.01 (1.00)	−0.18 (0.97)	−0.37 (0.95)	0.181	0.080	.029
Reading Span: Total Score Ascending	0.00 (1.06)	0.00 (0.98)	−0.13 (0.94)	−0.24 (1.00)	0.108	0.082	.186
Episodic memory					0.222	0.054	<.001
CVLT Total of Trials 1–5	−0.10 (1.00)	0.03 (1.00)	0.04 (1.03)	−0.23 (1.02)	0.236	0.085	.007
CVLT Short Delay Free Recall	−0.09 (0.99)	0.02 (1.01)	−0.01 (1.05)	−0.21 (1.12)	0.186	0.092	.039
CVLT Long Delay Free Recall	−0.11 (1.02)	0.02 (0.99)	0.04 (1.03)	−0.11 (1.04)	0.125	0.086	.158
WMS-III Logical Memory: Immediate Recall Story Units Total Score	−0.07 (1.00)	0.01 (0.99)	−0.07 (1.01)	−0.37 (0.95)	0.293	0.079	<.001
WMS-III Logical Memory: Delayed Recall Story Units Total Score	−0.09 (1.04)	0.02 (0.99)	−0.11 (1.03)	−0.39 (1.00)	0.256	0.083	.003
WMS-III Visual Reproduction: Immediate Recall Total Score	−0.16 (1.01)	0.03 (1.00)	0.04 (1.02)	−0.16 (0.98)	0.166	0.082	.055
WMS-III Visual Reproduction: Delayed Recall Total Score	−0.17 (1.04)	0.03 (0.99)	0.12 (0.99)	−0.21 (0.93)	0.295	0.078	<.001
Processing speed					0.197	0.057	.002
D-KEFS Trails 2 time (log transformed)	−0.07 (1.00)	0.01 (1.00)	0.03 (1.05)	−0.2 (1.03)	0.213	0.086	.018
D-KEFS Trails 3 time (log transformed)	−0.12 (1.04)	0.03 (1.01)	−0.01 (1.01)	−0.26 (1.04)	0.223	0.086	.010
Stroop: Raw Word Score [46]	0.06 (1.04)	−0.01 (1.00)	−0.31 (0.99)	−0.44 (0.92)	0.148	0.076	.076
Stroop: Raw Color Score [46]	0 (1.02)	−0.01 (0.99)	−0.29 (0.98)	−0.49 (0.91)	0.202	0.075	.015
Abstract Reasoning					0.159	0.087	.066
WASI Matrix Reasoning Raw Score	−0.11 (0.96)	0.02 (1.00)	−0.03 (1.01)	−0.21 (1.05)	0.159	0.087	.066
Verbal fluency					0.085	0.057	.254
D-KEFS Letter Fluency FAS total correct	0.05 (0.98)	0.00 (1.00)	−0.11 (0.99)	−0.18 (1.09)	0.088	0.088	.301
D-KEFS Category Fluency Animal/Boys total correct	−0.04 (1.02)	0.01 (1.01)	−0.01 (1.02)	−0.11 (1.04)	0.082	0.086	.338

Abbreviations: D-KEFS, Delis-Kaplan Executive Function System [47]; CVLT, California Verbal Learning Test [48]; WASI, Wechsler Abbreviated Scale of Intelligence [49]; WMS-III, Wechsler Memory Scale-III [50].

NOTE. All scores have been adjusted by regressing out age 20 GCA scores and then converted to z-scores based on wave 1 means and standard deviations (SDs). Standard errors (SEs) were calculated using bootstrap resampling, and *P* values were calculated using permutation testing. Practice effect values represent the expected gain in performance resulting from repeat testing. These values can be subtracted from returnees' time 2 scores to adjust follow-up tests for the effects of practice.

Practice effects for each resample were then recalculated, and the standard error was calculated as the SD across resamples.

Significance of each practice effect was estimated using permutation testing (10,000 permutations). At each permutation, the difference score was calculated after randomly shuffling labels of returnees and ARs at wave 2; the attrition effect was calculated after randomly swapping the labels of returnees and dropouts at wave 1. The two-sided *P* value was calculated as the percentage of permutations in which the absolute value of the permuted practice effect was larger than the observed practice effect.

2.6. *Effect of adjusting for practice effects on analysis of change*

To examine whether practice effects potentially mask change in performance across time, we tested for change with and without adjustment. Change in performance of returnees between wave 1 and wave 2 was tested with paired t-tests.

2.7. *Effect of adjusting for practice effects on classification of MCI*

Classification of returnees with MCI was compared both before and after correcting for practice effects. We defined

MCI according to the Jak-Bondi approach as described previously [25,26]. Neuropsychological measures were adjusted for early adult GCA so that MCI would reflect change over time rather than just longstanding low cognitive performance. Impairment was defined as having 2+ measures within a domain >1.5 SD below age- and education-adjusted normative means. Individuals with an impaired memory domain were further specified as amnestic mild cognitive impairment (aMCI) patients, and those with impairments in domains other than memory were classified as non-aMCI patients. Differences in the proportion of individuals classified as having MCI at wave 2 before and after adjusting for practice effects were assessed with McNemar's χ^2 test.

3. Results

3.1. Practice effects after 6 years

The following results are based on scores adjusted for age 20 GCA (see Supplementary Table S1 for unadjusted and education-adjusted results). The magnitude of practice effects ranged from 0.082 to 0.388 SD units (Table 2). Of the 25 individual measures, there were significant ($P < .05$) practice effects for 14 measures, with an additional four measures showing trend-level effects ($.05 < P < .1$). There were significant practice effects for four of the seven cognitive abilities: visual-spatial, short-term/working memory, episodic memory, and processing speed (Fig. 1). Unadjusted and education-adjusted practice effects were significant for all seven cognitive abilities (Fig. 1). Attrition effects were minimal, with only a few tests demonstrating significant effects (see Supplementary Table S2). The magnitude of significant attrition effects were small, ranging from 0.025 to 0.036 SD units.

3.2. Impact on classification of MCI

Among returnees, 106 individuals (11%) were classified as having any MCI at wave 1. At wave 2, 90 participants (9.3%) were classified as having any MCI. However, after correcting for practice effects adjusted for age 20 GCA, 147 participants (15.2%) were classified as having any MCI (see Fig. 2). Among returnees, this represents almost a doubling in the rate of individuals who converted to MCI, from 4.5% to 9%, a substantial increase (McNemar's $\chi^2(1) = 55.02$, $P < .001$). The rate of reversion from any MCI to cognitively normal among returnees also decreased from 6.1% to 4.8%.

These results can further be broken down by MCI subtype. At wave 1, 68 participants (7%) were classified as aMCI patients. At wave 2, 60 participants (6.2%) were classified as aMCI patients before correcting for practice effects, and 91 participants (9.4%) were classified as aMCI patients after correction. This was a significant increase (McNemar's $\chi^2(1) = 29.03$, $P < .001$). There were 38 participants classified as non-aMCI patients at wave 1 (4%). At wave 2, 30 participants (3%) were classified as non-aMCI patients before correction and 56 participants (5.8%) after correction. This was also a significant increase (McNemar's $\chi^2(1) = 20.83$, $P < .001$).

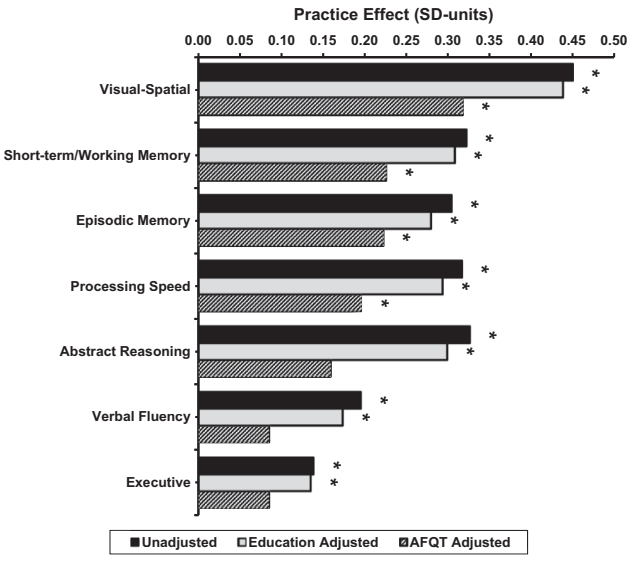

Fig. 1. Practice effects by cognitive domain. Practice effects are presented in standard deviation (SD) units for each cognitive ability composite. Bars represent practice effects calculated from unadjusted scores, scores adjusted for education, and scores adjusted for age 20 AFQT. Asterisks indicate significant practice effects ($P < .05$) as determined by permutation testing. Unadjusted practice effects are likely to be biased upward due to significant group differences in age 20 AFQT and education, thus adjusted scores reflect more accurate estimates. Abbreviation: AFQT, Armed Forces Qualification Test.

3.3. Impact on change in cognitive measures

Table 3 shows results using the typical approach of assessing change in returnees only, without replacement participants or correcting for practice effects. Based on this analysis, there were significant declines in five of the seven cognitive domains: visual-spatial, short-term/working memory, verbal fluency, executive function, and processing speed. However, after adjusting for practice effects in the returnees, significant decreases were found in all seven domains. As we found previously, with no adjustment for practice effects, The California Verbal Learning Test-II remained relatively stable ($t_{(976)} = -0.206$, $P = .837$), there was a significant decrease in Logical Memory ($t_{(985)} = -3.969$, $P < .001$), and Visual Reproduction showed a significant increase ($t_{(986)} = 1.965$, $P = .050$). After adjustment, however, all three episodic memory tests demonstrated significant decreases (California Verbal Learning Test-II: $t_{(976)} = -6.101$, $P < .001$; Logical Memory: $t_{(985)} = -14.679$, $P < .001$; and Visual Reproduction: $t_{(986)} = -7.166$, $P < .001$; Fig. 3).

4. Discussion

4.1. Under diagnosis of MCI

The results from the present study demonstrate that significant practice effects are present across most cognitive

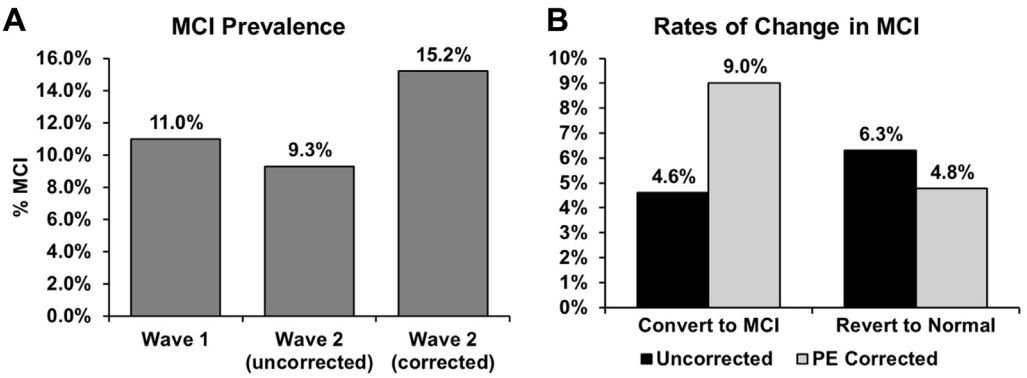

Fig. 2. Impact of practice effect adjustment on diagnosis of MCI. (A) Rate of MCI at each time point before and after correction for PEs. (B) Rates of change in MCI from wave 1 to wave 2 before and after correction for PEs. Abbreviations: MCI, mild cognitive impairment; PE, practice effect.

domains even after a 6-year interval between test administrations from an average age of 56 to 62 years. Importantly, there was a substantial impact of practice effect adjustment on the identification of MCI. The conversion rate increased while reversion from MCI to normal decreased, representing a substantial underdiagnosis of impairment at follow-up testing when practice effects were not taken into account. This finding has serious implications for clinical practice, drug intervention trials, general cognitive aging research, and research on AD. MCI diagnoses often rely on comparisons to age-adjusted norms, which assume scores result from an initial test administration or at least do not contain practice effects [26,27]. Our results strongly suggest that if follow-up test scores are not adjusted for practice effects, a meaningful number of individuals who have actually converted to MCI will be missed and might not be detected until the subsequent follow-up assessment when more substantial decline has occurred.

There is ample support for the validity of the Jak-Bondi approach to MCI diagnosis [25,41]. In the VETSA sample, we previously found that higher Alzheimer's polygenic risk scores were associated with significantly increased odds of having Jak-Bondi–defined MCI [42]. However, it is worth noting that these results are applicable to all definitions of MCI because all diagnostic approaches include assessment of cognitive decline. The National Institute on Aging–Alzheimer's Association criteria [43] define impairment as lower performance "than would be expected for the patient's age and educational background." They also state that "If repeated assessments are available, then a decline in performance should be evident over time." Our findings should thus be entirely applicable to this widely accepted definition of MCI.

AD prevention trials have largely been unsuccessful, and a commonly proposed explanation is that treatment is initiated too late in the disease course [28]. These trials require multiple administrations of cognitive tests and are thus subject to contamination by practice effects. Detecting the earliest stages of impairment is critical not only for identifying individuals most in need of treatment but also in developing cognitive outcome measures that are sensitive to change. These adjustments are most important for early identification, that is, at the earliest stages of decline when individuals are close to the cutoffs for impairment. Individuals with a diagnosis of AD are already below threshold, and a previous study found that adjusting for practice effects did not improve discrimination between normal controls and AD patients over baseline scores [29]. This is likely because at later disease stages, the magnitude of impairment is very large relative to any practice-related gains. The extent to which practice effects change across disease stages is also likely to be

Table 3
Change in cognitive performance

Cognitive ability	Wave 1	Wave 2 unadjusted	Wave 2 adjusted	$t_{unadjusted}$	$P_{unadjusted}$	$t_{adjusted}$	$P_{adjusted}$
Visual-spatial	−0.010 (0.81)	−0.107 (0.80)	−0.425 (0.80)	5.932	**<.001**	24.876	**<.001**
Short-term/working memory	0.011 (0.64)	−0.114 (0.62)	−0.339 (0.62)	9.511	**<.001**	26.404	**<.001**
Episodic memory	0.022 (0.69)	0.008 (0.72)	−0.214 (0.72)	0.918	.359	14.348	**<.001**
Abstract reasoning*	0.021 (1.0)	−0.027 (1.01)	−0.186 (1.01)	1.782	.075	7.322	**<.001**
Verbal fluency	0.005 (0.87)	−0.060 (0.87)	−0.145 (0.87)	3.363	**.001**	7.747	**<.001**
Processing speed	0.003 (0.73)	−0.151 (0.77)	−0.347 (0.77)	9.749	**<.001**	21.836	**<.001**
Executive	0.011 (0.635)	−0.129 (0.66)	−0.214 (0.66)	6.117	**<.001**	9.911	**<.001**

NOTE. Wave 2 is presented both unadjusted and adjusted for practice effects. Results from paired t-tests between waves 1 and 2 are presented. Means and standard deviations (SDs) of cognitive abilities at each wave for returnees only. Significant values are in bold.

*Indicates cognitive ability score based on a single measure; and other scores are composites of multiple test measures.

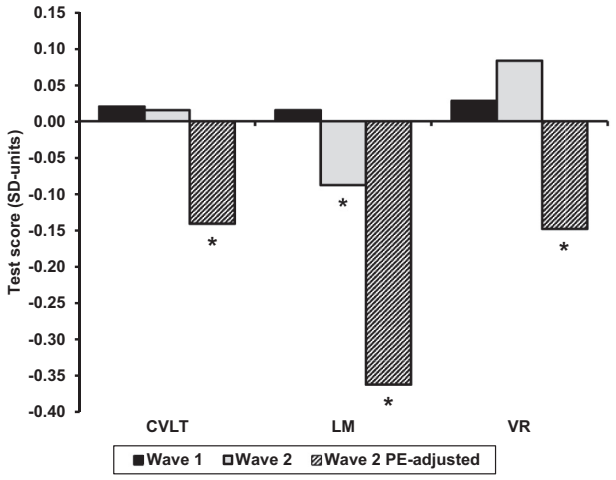

Fig. 3. Impact of PEs on individual episodic memory tests. Individual test score composites of the Episodic Memory ability at wave 1 and wave 2. Before correcting for PEs, the scores on the CVLT showed little change, scores on Logical Memory went down, and scores on Visual Reproduction went up. After PE adjustment, all tests showed decreases at wave 2. *Indicates significant ($P < .001$) difference compared to wave 1 score. Abbreviations: CVLT, California Verbal Learning Test; PE, practice effect.

test-specific; therefore, future studies are needed to examine these differences. Based on the present study results, we would also put forth the strong recommendation that the inclusion of replacement subjects ought to be a basic design feature of studies assessing cognition with respect to risk for MCI or AD, as well as other longitudinal cognitive aging studies.

4.2. Potential mechanisms of practice effects

Various mechanisms underlying the practice effect have been proposed throughout the literature, and it is likely that several of these explanations drive the results found here. For instance, participants may explicitly recall specific items from previous administrations on memory tests. Anecdotally, we have observed that some individuals do recall some aspects of the gist of stories from the previous administration (e.g., "the one about the woman who got robbed"). However, practice effects are found even when alternate forms are used [30]. It is also virtually impossible that individuals explicitly recalled items on tests such as Digit Span. Rather, these effects likely manifest due to increased comfort with the testing scenario, procedural memory, and/or learning new test-taking strategies [14,31]. The finding that performance gains are maintained over 6 years is consistent with a previous study, which estimated that it takes at least 7 years for practice effects to disappear [12]. Likewise, results from a recent meta-analysis [8] would predict that a 56-year-old retested after 5.6 years on an auditory attention/working memory study should show −0.147 SD change. This is similar to the −0.157 SD change among returnees on the Digit Span forward test after correcting for practice effects.

4.3. Practice effects when scores do not improve

These findings indicate that in samples expected to show normative declines in performance such as studies of aging, MCI, and AD, it is not sufficient to define practice effects only as performance increases. We observed decreases in performance over time before practice effect correction. This can be explained by the presence of normative (age-related) declines that are of a similar or greater magnitude than the practice effects. By incorporating attrition replacements, it is possible to decompose these effects. For example, returnees at time 1 and attrition replacements at time 2 are both taking the tests for the first time. The key difference is that attrition replacements are older. If attrition effects are minimal, we can interpret this comparison as representing age-related change. In contrast, returnees and attrition replacements at time 2 are both of similar age, and the key difference is that returnees have taken the tests before. This comparison roughly yields the expected practice effect (again, assuming minimal attrition effects). In examining Table 2, we find that in many cases, the age-related decline is expected to be as large as or larger than expected practice effects, resulting in an apparent decline in performance on many tests. Under a standard definition, this would be considered evidence of no practice effects. Yet, calculating practice effects through the use of attrition replacements demonstrates that practice effects were simply masking the true extent of decline. Put another way, a decline can still represent a practice effect if it is of smaller magnitude than the expected or normative decline.

An additional consideration is that returnees at time 1 may also be different from replacements at time 2 because the former are, in effect, selected for continued participation. This subgroup is typically better performing at time 1 than those who will not return. For this reason, the method also accounts for attrition effects to avoid inflating the practice effect estimates. These findings also support efforts to resolve differences between longitudinal and cross-sectional studies of cognitive aging. Longitudinal studies often find stable cognitive performance until around the age of 60 years, whereas cross-sectional studies find more gradual declines beginning earlier in life. Previous studies have found that correcting longitudinal studies for practice effects produces similar cognitive trajectories to those found in cross-sectional data [32,33]. Similarly, we found that scores on episodic memory tests, which remained stable or even increased, all demonstrated significant decline after adjusting for practice.

The results of this study underscore the importance of matching returnees and AR participants. The practice effect calculations could be inflated if differences were really due to differential premorbid ability between returnees and replacements. Also, there is evidence that individuals with higher IQs demonstrate larger practice-related gains [34].

They may learn more effective strategies after initial testing or better retain memory from previous administrations. To account for such differences, we adjusted all scores for age 20 GCA before calculating practice effects. Because studies rarely have access to early-life GCA, we also presented results for education-adjusted scores in the Supplementary Material. Our actual GCA measure provides a more accurate index of overall early adult cognitive ability than education and therefore more effectively addresses group differences in premorbid ability. We believe the education-adjusted estimates of practice effects are likely biased upward in our sample. However, adjustment for education has been used to calculate practice effects in prior studies [13], and earlier GCA measures are rarely available. Taken together, this set of results highlights the importance of proper matching in calculating practice effects.

4.4. Change scores versus practice effects

Our focus on assessing the impact of practice effects on the diagnosis of MCI should be differentiated from efforts to assess reliable change. Methods such as the Reliable Change Index (RCI) [35] or Standardized Regression-Based (SRB) [36] scores are focused on the analysis of change scores and are concerned primarily with individual differences in change. The method of Rönnlund et al. is meant to answer the question: what is the average increase in performance on a given test that we should expect simply due to having taken that test before? The focus is on the follow-up test performance rather than change per se. Assuming the attrition replacements are drawn from the same population, we can use their scores to answer this question because any differences can be assumed to be due to repeated exposure, that is, the average effect of practice on a given test for a given population. This difference is then adjusted for attrition effects so as not to inflate the practice effect because returning participants are often higher performing than non-returners. Now consider the example of older adults. For that population, change is essentially a function of both age-related decline (due to changes in physical health, brain health, etc.) and practice effects. The method of Rönnlund et al. makes it possible to disentangle these two effects, something that RCI and SRB score methods alone cannot do.

Another distinction centers on the fact that RCI and SRB score methods focus on relative change or individual differences in change. However, diagnosis is ultimately based on absolute scores. Regardless of how much a person may decline, they do not meet criteria for a diagnosis of MCI until their cognitive performance is below a particular cutoff. If a person is above the cutoff for diagnosis due to practice effects but would have been below the cutoff if they had taken the tests for the first time, it means that appropriate detection of the diagnosis will be delayed. The method of Rönnlund et al. is meant to produce follow-up scores that can independently be compared to normative thresholds, which aids in diagnosis. That comparison requires removing practice effects but retaining age-related decline. RCI and SRB score methods do not separate these two effects for the purposes of generating a follow-up score that can be compared to normative thresholds. The Rönnlund et al. approach also allows for cross-sectional analyses of time 2 data to combine both first-time test takers with returnees.

SRB change scores examine how performance changes compared with what is predicted given baseline characteristics of an individual. This approach is useful to identify individual-level factors that contribute to cognitive change. The RCI is used to assess whether a change score represents true change in performance above and beyond test-related error variance [37]. These methods are not in competition and can be used separately or in conjunction, with the Rönnlund et al. method serving as a precursor to more formal analyses of change scores. It is also important to note that the Rönnlund method subtracts a single value per test across all individuals and therefore does not remove any of the individual variability that is of interest in analyses of change. Therefore, RCI and SRB score approaches can be used after having estimated practice effects according to the method of Rönnlund et al.

4.5. Limitations

A limitation of the VETSA is that it is an all-male sample. Thus, these results may not generalize to women. There are mixed findings regarding whether there are sex effects on selective attrition [38–40]. Attrition effects were small relative to practice effects in the present study, and there were no sex differences reported in the mixed-sex study of Rönnlund et al. for attrition or practice effects [13]. Therefore, differential attrition effects based on sex are unlikely to have had a substantial impact on the estimated practice effects. The 5.6-year interval between assessments may be longer than other research studies or clinical exams, which may be collected every year or two. Although the exact values reported here may not be applicable to all studies, it should be emphasized that this approach can be used to calculate practice effects at any interval in any age group. Moreover, including replacement participants who are specific to each particular study will yield the most precise practice effect estimates. Our finding of significant practice effects at this longer interval also underscores the critical need to account for practice effects, which are likely to have increased impact at shorter intervals. The approach used here only accounts for a single follow-up session, but it

could also be scaled to situations with multiple follow-up sessions in which individuals differ in the number of follow-ups.

4.6. Summary

In sum, the results of the present study strongly suggest the importance of correcting for practice effects in longitudinal studies of older adults, particularly those focused on detecting subtle and early signs of cognitive impairment. Alternative methods of accounting for these effects have been proposed, such as the dual-baseline approach [4] and alternate testing forms [8]. Although these strategies have been shown to attenuate practice effects somewhat, performance gains occur even after the third and fourth sessions [10,34], and when alternate forms are used [15]. Using attrition replacements does appear to be the most accurate method of estimating practice effects. Future work will attempt to extend these methods to assess more than two follow-up time points. More work is also necessary to further explore how practice effects change across the lifespan and how test-specific practice effects differ between groups in the early stages of various types of dementia. Nevertheless, the age of the present study cohort is particularly relevant for early detection of risk for AD and related dementias. Indeed, our results indicate that failure to properly account for practice effects can result in substantial underestimation and, hence, delayed detection of MCI.

RESEARCH IN CONTEXT

1. Systematic review: We searched PubMed for literature on practice effects. Almost all studies assume that only improved scores indicate practice effects. These studies have not focused on the impact of practice effects on early identification of mild cognitive impairment (MCI). Relevant references are appropriately cited.
2. Interpretation: By comparison with replacement participants, this study showed significant practice effects—despite mean-level declines—across multiple cognitive domains after a 6-year interval. Correcting for practice effects resulted in a doubling of the conversion rate to MCI, suggesting that not accounting for practice effects may substantially hinder efforts to identify the earliest stages of Alzheimer's disease progression.
3. Future directions: These findings underscore the need to account for practice effects. We strongly recommend that longitudinal study designs incorporate replacement participants to gauge practice effects. Further work is needed to assess the impact of practice effects on identification of cognitive decline across different ages.

Acknowledgments

The content of this article is the responsibility of the authors and does not represent official views of NIA/NIH, or the Veterans' Administration. Numerous organizations provided invaluable assistance in the conduct of the VET Registry, including U.S. Department of Veterans Affairs, Department of Defense; National Personnel Records Center, National Archives and Records Administration; Internal Revenue Service; National Opinion Research Center; National Research Council, National Academy of Sciences; and the Institute for Survey Research, Temple University. The authors gratefully acknowledge the continued cooperation of the twins and the efforts of many staff members. The study was supported by awards from the National Institutes of Health/National Institute on Aging (R01s AG018386, AG022381, AG022982, and AG050595 to W.S.K.; R01 AG018384 to M.J.L.; R03 AG 046413 to C.E.F.; and K08 AG047903 to M.S.P).

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.dadm.2018.04.003.

References

[1] Nesselroade JR, Baltes PB. Longitudinal Research in the Study of Behavior and Development. New York: Academic Press; 1979.
[2] Schaie KW. The course of adult intellectual development. Am Psychol 1994;49:304–13.
[3] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–92.
[4] McCaffrey RJ, Westervelt HJ. Issues associated with repeated neuropsychological assessments. Neuropsychol Rev 1995; 5:203–21.
[5] Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. Clin Neuropsychol 2010;24:1267–78.
[6] Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimers Dement (Amst) 2015;1:103–11.
[7] Mathews M, Abner E, Kryscio R, Jicha G, Cooper G, Smith C, et al. Diagnostic accuracy and practice effects in the National Alzheimer's Coordinating Center Uniform Data Set neuropsychological battery. Alzheimers Dement 2014;10:675–83.

[8] Calamia M, Markon K, Tranel D. Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. Clin Neuropsychol 2012;26:543–70.
[9] Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, et al. Practice effects predict cognitive outcome in amnestic mild cognitive impairment. Am J Geriatr Psychiatry 2011;19:932–9.
[10] Bartels C, Wegrzyn M, Wiedl A, Ackermann V, Ehrenreich H. Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. BMC Neurosci 2010;11:118.
[11] Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc 2003; 9:419–28.
[12] Salthouse TA, Schroeder DH, Ferrer E. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. Dev Psychol 2004;40:813–22.
[13] Rönnlund M, Nyberg L, Backman L, Nilsson LG. Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. Psychol Aging 2005;20:3–18.
[14] Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology 2008;22:800–11.
[15] Hinton-Bayre A, Geffen G. Comparability, reliability, and practice effects on alternate forms of the Digit Symbol Substitution and Symbol Digit Modalities tests. Psychol Assess 2005;17:237–41.
[16] Schaie KW, Labouvie GV, Barrett TJ. Selective attrition effects in a fourteen-year study of adult intelligence. J Gerontol 1973;28:328–34.
[17] Panizzon MS, Neale MC, Docherty AR, Franz CE, Jacobson KC, Toomey R, et al. Genetic and environmental architecture of changes in episodic memory from middle to late middle age. Psychol Aging 2015;30:286–300.
[18] Kremen WS, Franz CE, Lyons MJ. VETSA: The Vietnam Era Twin Study of Aging. Twin Res Hum Genet 2013;16:399–402.
[19] Schoeneborn CA, Heyman KM. Health Characteristics of Adults Aged 55 Years and Over: United States, 2004–2007. National Health Statistics Reports; no. 16. National Health Statistics Reports. Hyattsville, MD: National Center for Health Statistics; 2009.
[20] Eisen S, True W, Goldberg J, Henderson W, Robinette CD. The Vietnam Era Twin (VET) Registry: Method of construction. Acta Genet Med Gemellol (Roma) 1987;36:61–6.
[21] Henderson WG, Eisen S, Goldberg J, True WR, Barnes JE, Vitek ME. The Vietnam Era Twin Registry: A resource for medical research. Public Health Rep 1990;105:368–73.
[22] Bayroff AG, Anderson AA. Development of Armed Forces Qualification Test 7 and 8 (Technical Research Report 1122). Alexandria, Virginia: US Army Research Institute; 1963.
[23] Lyons MJ, York TP, Franz CE, Grant MD, Eaves LJ, Jacobson KC, et al. Genes determine stability and the environment determines change in cognitive ability during 35 years of adulthood. Psychol Sci 2009;20:1146–52.
[24] Vuoksimaa E, Panizzon MS, Chen CH, Eyler LT, Fennema-Notestine C, Fiecas MJ, et al. Cognitive reserve moderates the association between hippocampal volume and episodic memory in middle age. Neuropsychologia 2013;51:1124–31.
[25] Bondi MW, Edmonds EC, Jak AJ, Clark LR, Delano-Wood L, McDonald CR, et al. Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. J Alzheimer's Dis 2014;42:275–89.
[26] Jak AJ, Bondi MW, Delano-Wood L, Wierenga C, Corey-Bloom J, Salmon DP, et al. Quantification of five neuropsychological approaches to defining mild cognitive impairment. Am J Geriatr Psychiatry 2009;17:368–75.
[27] Petersen RC. Mild cognitive impairment as a diagnostic entity. J Intern Med 2004;256:183–94.
[28] Golde TE, Schneider LS, Koo EH. Anti-abeta therapeutics in Alzheimer's disease: the need for a paradigm shift. Neuron 2011;69:203–13.
[29] Zehnder AE, Blasi S, Berres M, Spiegel R, Monsch AU. Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? Am J Alzheimers Dis Other Demen 2007;22:416–26.
[30] Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, et al. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol 2005; 20:517–29.
[31] Anastasi A, Urbina S. Psychology Testing. New Jersey: Prentice Hall; 1997.
[32] Rönnlund M, Nilsson LG. Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. Intelligence 2006; 34:63–78.
[33] Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology 2010;24:563–72.
[34] Rabbitt P, Lunn M, Wong D, Cobain M. Age and ability affect practice gains in longitudinal studies of cognitive change. J Gerontol B Psychol Sci Soc Sci 2008;63:P235–40.
[35] Heaton RK, Temkin N, Dikmen S, Avitable N, Taylor MJ, Marcotte TD, et al. Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. Arch Clin Neuropsychol 2001;16:75–91.
[36] McSweeny AJ, Naugle RI, Chelune GJ, Lüders H. "TScores for Change": An illustration of a regression approach to depicting change in clinical neuropsychology. Clin Neuropsychol 1993;7:300–12.
[37] Blasi S, Zehnder AE, Taylor KI, Berres M, Spiegel R, Monsch AU. Norms for Change in Episodic Memory as a Prerequisite for the Diagnosis of Mild Cognitive Impairment (MCI). Neuropsychology 2009;23:189–200.
[38] Rabbitt P, Watson P, Donlan C, Bent N, McInnis L. Subject attrition in a longitudinal study of cognitive performance in community-based elderly people. Facts Res Gerontol 1994;14:203–7.
[39] Salthouse TA. Selectivity of attrition in longitudinal studies of cognitive functioning. J Gerontol B Psychol Sci Soc Sci 2014;69:567–74.
[40] Van Beijsterveldt CE, van Boxtel MP, Bosma H, Houx PJ, Buntinx F, Jolles J. Predictors of attrition in a longitudinal cognitive aging study: the Maastricht Aging Study (MAAS). J Clin Epidemiol 2002;55:216–23.
[41] Jak AJ, Panizzon MS, Spoon KM, Fennema-Notestine C, Franz CE, Thompson WK, et al. Hippocampal atrophy varies by neuropsychologically defined MCI among men in their 50s. Am J Geriatr Psychiatry 2015;23:456–65.
[42] Logue MW, Panizzon MS, Elman JA, Gillespie NA, Hatton SN, Gustavson DE, et al. Use of an Alzheimer's disease polygenic risk score to identify mild cognitive impairment in adults in their 50s [published online ahead of print February 27, 2018]. Mol Psychiatry 2018; https://doi.org/10.1038/s41380-018-0030-8.
[43] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9.
[44] Ekstrom RB, French JW, Harman HH, Dermen D. Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational testing service; 1976.
[45] Thurstone LL. A Factorial Study of Perception. Chicago, IL: University of Chicago Press; 1944.
[46] Golden CJ. Stroop Color and Word Test: A Manual for Clinical and Experimental Uses. Chicago, IL: Stoelting; 1978.
[47] Delis DC, Kaplan E, Kramer JH. Delis-Kaplan Executive Function System: Technical Manual. San Antonio, Texas: Psychological Corporation; 2001.
[48] Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test. 2nd ed. San Antonio, TX: Psychological Corporation; 2000.
[49] Wechsler D. Manual for the Wechsler Abbreviated Intelligence Scale (WASI). San Antonio, TX: The Psychological Corporation; 1999.
[50] Wechsler D. Wechsler Adult Intelligence Scale-III (WAIS-III) manual. San Antonio, TX: The Psychological Corporation; 1997.