Practice effects persist over two decades of cognitive testing: Implications for longitudinal research

Jeremy A. Elman[a,b], Erik Buchholz[a,b], Rouhui Chen[c], Mark Sanderson-Cimino[d], Tyler R. Bell[a,b], Nathan Whitsel[a,b], Katherine J. Bangen[a,e], Alice Cronin-Golomb[f], Anders M. Dale[f,g], Lisa T. Eyler[a], Christine Fennema-Notestine[a,b,g], Carol E. Franz[a,b], Nathan A. Gillespie[i], Eric L. Granholm[a], Daniel E. Gustavson[j], Donald J. Hagler Jr.[g], Richard L. Hauger[a,b,k], Diane M. Jacobs[h], Amy J. Jak[a], Mark W. Logue[l,m,n,o], Michael J. Lyons[f], Ruth E. McKenzie[p], Michael C. Neale[i], Robert A. Rissman[q], Chandra A. Reynolds[j], Rosemary Toomey[f], Arthur Wingfield[r], Hong Xian[s,t], Xin M. Tu[u], William S. Kremen[a,b,*], Matthew S. Panizzon[a,b,*]

* Joint senior authors
[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] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Evanston, IL, USA
[d] Memory and Aging Center, Weill Institute for Neurosciences, San Francisco, CA, USA,
[e] VA San Diego Healthcare System, San Diego, CA, USA
[f] Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA
[g] J. Craig Venter Institute, La Jolla, CA, USA
[h] Department of Neurosciences, University of California San Diego, La Jolla, CA, USA
[i] Virginia Institute for Psychiatric and Behavior Genetics, Richmond, Virginia, USA.
[j] Institute for Behavioral Genetics and Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, USA,
[k] Center of Excellence for Stress and Mental Health (CESAMH), VA San Diego Healthcare System, San Diego, CA, USA
[l] Department of Psychiatry, Boston University School of Medicine, Boston, MA, USA.
[m] Biomedical Genetics, Boston University School of Medicine, Boston, MA, USA.
[n] Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.
[o] Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA.
[p] School of Education and Social Policy, Applied Human Development and Community Studies, Merrimack College, North Andover, MA, USA
[q] Department of Physiology and Neuroscience, Alzheimer's Therapeutic Research Institute of the Keck School of Medicine of the University of Southern California, San Diego, CA, USA
[r] Department of Psychology and Volen National Center for Complex Systems, Brandeis University, Waltham, MA, USA
[s] Department of Epidemiology and Biostatistics, Saint. Louis University, St. Louis, Missouri, USA
[t] Research Service, VA St. Louis Healthcare System, St. Louis, Missouri, USA
[u] Herbert Wertheim School of Public Health & Human Longevity Science, University of California San Diego, La Jolla, CA, USA

Correspondence:
Jeremy A. Elman
Department of Psychiatry
University of California, San Diego
9500 Gilman Dr. (MC0738)
La Jolla, CA 92093
Phone: (858) 534-6842
E-mail: jaelman@health.ucsd.edu

ABSTRACT

Background: Repeated cognitive testing can boost scores due to practice effects (PEs), yet it remains unclear whether PEs persist across multiple follow-ups and long durations. We examined PEs across multiple assessments from midlife to old age in a nonclinical sample.

Method: Men (N=1,608) in the Vietnam Era Twin Study of Aging (VETSA) underwent neuropsychological assessment comprising 30 measures across 4 waves (~6-year testing intervals) spanning up to 20 years. We leveraged age-matched replacement participants to estimate PEs at each wave. We compared cognitive trajectories and MCI prevalence using unadjusted versus PE-adjusted scores.

Results: Across follow-ups, a range of 7-12 tests (out of 30) demonstrated significant PEs, especially in episodic memory and visuospatial domains. Adjusting for PEs resulted in improved detection of cognitive decline and MCI, with up to 20% higher MCI prevalence.

Conclusion: PEs persist across multiple assessments and decades underscoring the importance of accounting for PEs in longitudinal studies.

Keywords: practice effects; repeat testing; serial testing; longitudinal testing; mild cognitive impairment; cognitive change

1. INTRODUCTION

Longitudinal designs are critical for understanding cognitive development and decline [1-4]. In the context of studies on Alzheimer's disease (AD) and dementia, longitudinal testing is important for identifying risk factors for cognitive decline, understanding variation in disease progression, and evaluating efficacy of treatments [5,6]. However, it has long been acknowledged that performance at follow-up may be artifactually inflated due to practice effects (PEs) [7-9]. PEs can obscure the true nature of cognitive decline at later ages with implications for AD and related dementia clinical trials that assess slowing of cognitive decline to determine treatment effects [10-14].

Despite widespread acknowledgment of PEs, accounting for them is not standard practice in many longitudinal studies. We previously demonstrated that PEs persisted across a 6-year interval, and that not accounting for PEs resulted in underdiagnosis of mild cognitive impairment (MCI) and a greater proportion of individuals reverting to normal cognitive status [15]. In an independent sample, we found that adjusting for PEs not only reduced rates of reversion to normal cognitive status, but most importantly, it led to detection of progression to MCI one year earlier [16,17]. Moreover, practice-adjusted diagnoses were validated based on greater concordance with AD biomarker positivity. We and others have also shown that adjusting for PEs can reduce sample sizes needed in clinical trials, resulting in substantial cost savings [13,16].

PEs are sometimes defined as improved performance at follow-up compared to baseline [18,19]. Even if there is observed decline, however, PEs may still be present and contribute to scores. In the context of older age or neurodegenerative disease when normative declines are expected (with higher magnitude than PEs), the observed impact of PEs may simply be attenuated decline [20]. One solution proposed by Ronnlund et al. [12] is to use attrition replacement (AR) participants, new participants recruited during follow-up study waves who are age-matched to the ongoing longitudinal cohort [see also 15-17,20,21]. By comparing well-

matched individuals who differ only on whether they have previously been tested, we can estimate the expected improvement in performance due to practice, even when there is an observed decrease in scores at follow-up. This method allows us to create an adjusted follow-up score that can be compared to norms or thresholds for impairment, which assume that test scores are driven by current ability and not practice.

Studies of PEs have typically examined test-retest intervals of less than 5 years, with most being in the range of 6 months to 2 years [22,23], and there has been little investigation of PEs in cohorts that have followed individuals for extended periods of time (i.e., over 10 years, but see [10] and [24] for notable exceptions). Extended longitudinal follow-ups within individuals are necessary to clarify the evolution of cognition across critical transition periods such as midlife to old age, yet PEs hinder our ability to accurately measure these trajectories. The Vietnam Era Twin Study of Aging (VETSA) presents a rare opportunity to examine PEs on cognitive performance in ~1,600 individuals assessed over 4 study waves and 20 years. We extend the Ronnlund et al. [12] method previously applied to two waves of VETSA data [15], given that the number of assessments has increased with varying patterns of missingness. Here, we apply a generalized approach that leverages AR participants and flexibly handles complex testing schedules to examine how PEs evolve over two decades, and the impact of adjusting for these effects on cognitive trajectories and rates of MCI. We expected that PEs on individual measures would vary by cognitive domain, with greatest effect for tests in the episodic memory domain, and that effects would wane over time. In addition, we compared trajectories of cognitive composite scores and MCI prevalence using unadjusted versus PE-adjusted scores. We predicted that adjusting for PEs would result in higher prevalence of MCI at each wave due to earlier detection of impairment.

2. METHODS

2.1. Participants

Participants were 1,608 individuals tested at one or more of four completed study waves of VETSA. VETSA is an on-going longitudinal study of aging and risk for AD beginning in middle age [25-27]. Participants were members of the Vietnam Era Twin (VET) Registry, a national, community-dwelling sample of male-male twins who served in the U.S. military during the Vietnam era (1965-1975) [28-30]. All Registry members were invited to participate in the Harvard Drug Study [31], for which ascertainment was not based on any diagnostic or substance use criteria. VETSA participants were then randomly recruited from the Harvard Drug Study sample. At baseline VETSA participants were similar to American men in their age cohort with respect to health, education, and lifestyle characteristics based on Center for Disease Control and Prevention data [32], and nearly 80% reported no combat exposure [30]. Wave 1 data collection occurred between 2003-2007 with follow-up data collections at wave 2 (2009-2014), wave 3 (2016-2019), and wave 4 (2021-2024).

Of the 1,608 individuals in the current study, 1,291 underwent wave 1 baseline assessments. At waves 2 and 3, attrition replacement participants age-matched to the ongoing sample were recruited from the VET Registry and tested for the first time (wave 2 n=193, wave 3 n=124). These participants were then invited for follow-up at all subsequent waves (see **Figure 1A** for all patterns of assessments). On average, participants were 56 years of age (range 51-61) at wave 1, 62 years (range 56-67) at wave 2, 68 years (range 61-73) at wave 3, and 74 (range 67-79) at wave 4 (see **Figure 1B** for age distributions by wave). The average time between wave 1 and 2 was 5.7 years, with 5.9 years between waves 2 and 3, and 5.6 years between waves 3 and 4.

Participants traveled to the University of California San Diego (UCSD) or Boston University (BU) to participate in VETSA. In some cases, when a participant was unable to travel to San Diego or Boston, examiners travelled to administer the same testing protocol at a site nearby the participant's place of residence. The study was performed in accordance with the ethical standards per the 1964 Declaration of Helsinki and later amendments. Informed consent was

obtained from all participants and institutional review boards at both sites approved all study procedures.

2.2. Cognitive tests and measures

A set of 30 neuropsychological measures covering multiple cognitive domains were assessed for PEs (see **Table 2** for full list of tests and scores by domain). Practice effect estimation and adjustment used raw scores (i.e., prior to any conversion to normed scores for purposes of classifying cognitive impairment). For consistency, scores from tests where lower values indicate better performance (e.g., reaction time) were reverse-coded so that higher scores uniformly indicate better performance. This transformation did not alter the underlying rank order or variance of the original scores and so had no impact on the shape of the distributions. In addition to the specific abilities, general cognitive ability (GCA) was measured with the validated Armed Forces Qualification Test (AFQT) [33]. The AFQT is a 100-item multiple-choice paper-and-pencil test administered during military induction at average age 20 and was again administered at each VETSA assessment. The AFQT is highly correlated with standard IQ measures (r=.84) [34]. AFQT scores from military induction (hereafter referred to as "age 20 AFQT") were reported as percentile scores. Therefore, they have been transformed to the standard normal distribution using a probit transform.

2.3. Estimation of practice effects

Practice effects were estimated for each of the 30 measures using generalized estimating equations (GEE). Estimation was based on raw scores (i.e., not normed). We opted for this class of semiparametric regression models because it requires no assumptions about data distributions such as normality to provide valid inferences for virtually all data distributions arising in practice [35]. Let Y_{iaw} denote the score on a cognitive test measure (dependent variable) and X_{iaw} the collection of all explanatory (independent) variables available for the ath

assessment of subject *i* occurring at wave *w.* As all cognitive outcomes were treated as continuous, the identity link function was used in specifying GEE model [35]:

$$\begin{aligned}\mathrm{E}[\mathrm{Y}_{\mathrm{iaw}}|X_{iaw}] = \beta_0 + \beta_{age}(age) + \beta_{afqt}(age\ 20\ AFQT) + \beta_{s1}\mathrm{I}(s=1) + \beta_{s2}\mathrm{I}(s=2) + \beta_1 I(w=2) \\ + \beta_2 I(w=3) + \beta_3 I(w=4) + \beta_4 I(w=2\ \&\ a=2) + \beta_5 I(w=3\ \&\ a=2) \\ + \beta_6 I(w=3\ \&\ a=3) + \beta_7 I(w=4\ \&\ a=3) + \beta_8 I(w=4\ \&\ a=4)\end{aligned}$$

Here, $I(\cdot)$ denotes an indicator function that takes the value 1 if the condition inside is true, and 0 otherwise. Wave *w* and Assessment *a* can take values from 1 to 4, with the condition that the value of assessment cannot be greater than the current study wave*.* Additional indicator variables capture the interaction between Wave and Assessment (again, with the condition that *a*≤*w*). Importantly, individuals at each wave can differ on how many previous assessments they participated in. Participants may also miss a study wave but then return for follow-up at subsequent waves. To account for the potential impact of longer intervals between assessments on practice effects, we include the Skip variable *s*. Skip *s* can take the values 1 or 2 to indicate whether they missed 1 or 2 waves prior to the current assessment.

We additionally adjusted for age and age 20 AFQT score. The age term allows us to estimate practice effects independent of normative age-related decline. We previously found that the attrition replacement group at wave 2 had significantly lower age 20 AFQT scores than the individuals enrolled at wave 1. Given that participants were randomly recruited from the same population (i.e., the VET Registry), this is likely due to random sampling variation. However, differences in “premorbid” cognitive ability between groups may artificially inflate PEs. For example, higher performance in the returnees at follow-up could be due to PEs, but could also be due to their higher average young adult GCA.

The coefficients from the GEE models described above can be interpreted as the expected difference in score on a particular measure associated with the relevant wave, assessment number, and whether any prior waves were skipped (see **Supplemental Table S1** for interpretation of model terms). Of greatest interest are coefficients β_3- β_8, which can be

interpreted as PE estimates: the expected difference in performance on a test at a given wave based on having taken that test one or more times in the past. The interpretation of β_3 is somewhat different than β_1 and β_2 despite not having an indicator for Assessment *a>1*. Because no replacement participants were recruited at wave 4, the reference level corresponds to a second assessment at wave 4 (*w=4 & a=2)*. Therefore, this coefficient can be interpreted as the PE for individuals who have participated in one assessment prior to wave 4 compared to individuals taking the test for the first time, conditional on age. Our primary interest is the patterns of PEs among individuals who completed all 4 assessment waves.

2.4. Adjusting test scores for practice effects

For a given individual, we can calculate the expected PE for a test score by summing up the coefficients corresponding to the wave and how many previous assessments they participated in, as well as whether they have skipped any previous assessments. The resulting "adjustment value" is then subtracted from their observed score, resulting in the score that we would expect if they had been taking the test for the first time. As an example, consider an individual who has completed all 4 waves of testing. We can adjust their score on, say, Logical Memory at wave 4 by subtracting the value of β_8 (i.e., the PE for someone completing a 4th assessment at wave 4). If they had missed wave 3, we would calculate the appropriate adjustment value by adding the estimated effect of skipping one prior assessment (i.e., β_{s1}) to the practice effect β_8. We might expect the adjustment value in this latter case to be smaller, reflecting a reduced PE due to a longer follow-up interval.

2.5. Effect of adjusting for practice effects on cognitive factor scores

We have previously shown that composite measures such as cognitive factor scores can improve reliability and prediction of cognitive decline [36]. Therefore, we are interested in determining the extent to which PEs in individual scores impact these composite measures of cognition. We calculated cognitive factor scores across 6 cognitive domains: episodic memory, executive function, verbal fluency, processing speed, visual memory and visuospatial ability.

Details of each factor model are described in previous publications [36-39]. The weights used to generate these factor scores were based on previous latent variable analyses of multiple neuropsychological tests. Two versions of each score were created, one using PE-adjusted scores and the other using unadjusted scores. Higher scores reflect better performance in each domain. To test for differences in PE-adjusted versus unadjusted cognitive composite scores at each wave, we used longitudinal GEE models with an interaction between wave and adjustment (adjusted versus unadjusted).

2.6. Effect of adjusting for practice effects on classification of MCI

Classification of MCI was compared both before and after adjusting for PEs. We defined MCI according to the Jak/Bondi approach as described previously [40,41]. To ensure that MCI classification captured cognitive decline rather than lifelong low ability, neuropsychological scores were adjusted using early adult GCA (age 20 AFQT scores) as a covariate. Impairment was defined as having 2+ measures within a domain >1.5 SD below age-based normative means. Individuals with an impaired memory domain were further classified as amnestic MCI (aMCI), and those with impairments in domains other than memory were classified as non-amnestic MCI (naMCI). Differences in the proportion of individuals classified as having MCI at each wave before and after adjusting for PEs were assessed with McNemar's χ^2 test.

3. RESULTS

3.1. Practice effects on cognitive tests across four study waves.

We focus on reporting PEs for individuals who attended all four waves. This was the most common pattern of participation and allowed us to examine the evolution of PEs over the longest period of follow-up and greatest number of assessments. **Figure 2** presents PE estimates and confidence intervals of all measures across each of the three follow-up assessments. The PE estimates can be interpreted as the expected boost in performance individuals receive from having taken the test before, potentially multiple times, compared to

individuals of a similar age and young adult GCA taking the test for the first time. At the first follow-up (wave 2), 12 of 30 measures demonstrated significant PEs ranging in magnitude from 0.14 SD units to 0.34 SD units. The largest effects occurred for visuospatial ability and measures of verbal episodic and visual memory. At the second follow-up (wave 3), 8 of the 30 measures demonstrated significant PEs ranging from 0.15 SD units to 0.33 SD units. Similar to the first follow-up, practice primarily affected visual memory and visuospatial measures. At the third follow-up (wave 4), 7 of the 30 measures demonstrated significant practice effects ranging from 0.23 SD units to 0.35 SD units. PEs. Overall, the number of significant practice effects declined slightly over time, though memory and visuospatial tasks consistently showed the strongest effects. Additionally, tests that showed significant practice effects at later follow-ups typically, but not always, showed significant practice effects at earlier follow-ups.

3.2. Impact on cognitive factor scores.

PE-adjustment led to significantly lower scores across all follow-up waves for all domains (all ps < 0.05), indicating that unadjusted scores may overestimate cognitive performance in these domains. Consistent with results from individual test measures, decreases in performance were most notable in the domains of verbal episodic memory (ranging between -0.24 to -0.26 SD units) and visual memory (ranging between -0.11 to -0.32 SD units), and were weakest in the fluency domain (ranging between -0.04 to -0.07 SD units). See **Figure 3** and **Supplemental Table S2** for full results of comparisons across waves.

3.3. Impact on classification of MCI

We next examined the impact of adjusting for PEs on the rate of MCI at follow-up assessments (**Figure 4**). Adjusting for PEs significantly increased the estimated prevalence of MCI at wave 2 (12.3% vs. 15.6%, χ^2=4.821, p = 0.028) and wave 4 (16.3% vs. 21%, χ^2=5.402, p=0.020), with a non-significant upward trend at wave 3 (15.1% vs. 18.1%, χ^2=3.548, p=0.060). These increases suggest that failure to adjust for practice effects may mask clinically

meaningful cognitive decline, particularly in long-term follow-up. Consistent with findings that measures in the memory domains exhibited the strongest PEs, the increased rates of MCI were primarily due to increased prevalence of amnestic MCI. After PE adjustment, the rate of amnestic MCI was significantly higher at all follow-up waves (wave 2: 8% vs 11%, χ^2=5.213, p=0.022; wave 3: 10% vs 13%, χ^2=6.333, p=0.012; wave 4: 13% vs 17%, χ^2=4.546, p=0.033). In contrast, rates of non-amnestic MCI remained stable across waves regardless of adjustment (all ps>0.05), suggesting that practice effects may play a lesser role in these domains.

4. DISCUSSION

These results demonstrate the persistence of PEs on certain cognitive measures over extended periods of time and multiple assessments. Our findings are broadly consistent with other studies in that PEs were most apparent at the first follow-up and were weaker – but not absent – at later follow-ups [19,42-44]. The decreasing number of tests exhibiting significant practice effects at later follow-ups may be partially explained by smaller sample sizes at these waves. It should also be noted that most prior studies finding that PEs did not occur at follow-up examined assessment schedules with much shorter intervals and did not use AR-participant approaches. Additionally, performance on certain tests may have plateaued at later follow-ups – suggesting no additional PEs – because participants had reached ceiling. Although our 5-6 year testing interval is comparatively long, other studies have also found evidence of PEs after 7 years [10,12,22,45]. Persistent PEs were most notable for memory and visuospatial domains and rates of MCI were up to 29% higher after adjusting for PEs, driven primarily by increases in amnestic MCI.

One of the few studies to examine PEs for multiple assessments over a comparable period of time found that individuals continued to show practice-related improvements on an intelligence test up to the fourth wave [10,11]. The current study adds additional context by demonstrating the variability in PEs across specific cognitive domains and measures.

Consistent with prior studies, we found PEs were most apparent and most persistent in the memory domain, whereas language, attention/working memory, and speed showed the smallest effects [18,19,22,42,46]. This is particularly important given that memory tests such as Logical Memory are used to diagnose amnestic MCI, and our results underscore this point by showing a significant increase in rates of amnestic MCI after PE adjustment. We additionally found strong effects in the visuospatial domain, which is consistent with some studies [19] but not others [22,42]. This likely reflects variability across studies in the specific measures used to assess each domain, as certain tests may be more susceptible to factors such as ceiling effects [42].

An important aspect of the AR-participants approach used here is that we were able to estimate and adjust for PEs even with observed declines in performance. Although observed increases will occur when the boosts to performance resulting from practice are larger in magnitude than true decline, in cases where actual decline is of larger magnitude than the PEs, PEs may serve only to attenuate or obscure accelerated decline [10,20]. This scenario becomes more likely among samples in which greater normative decline is expected such as in cohorts of older individuals or those at risk for dementia. Indeed, some studies find a lack of PEs at older ages or in individuals with AD pathology [18,19,43,47]. However, it has been shown that when appropriate methods are used, individuals with MCI and dementia (including prominent memory impairments) do demonstrate benefits from practice [48,49]. We found that by wave 4, adjusting for PEs resulted in over a 20% increase in the rate of MCI, even though most individuals were showing declines in performance. This supports our prior findings that not accounting for PEs delays diagnosis of impairment [17,20]. Delayed detection of MCI has important implications for clinical trials that use conversion to impairment or rates of cognitive worsening as outcomes. Moreover, as shown elsewhere, adjusting for PEs reduces the required sample size resulting in multi-million dollar cost savings and allows for more accurate estimate of treatment effects [13,14,16]. Finally, earlier detection of impairment means that earlier intervention and possible greater slowing of disease progression is possible.

The approach we took in the current study estimated PEs at the group level and applied the same correction to individuals based on their test-retest schedule. Therefore, adjustment only accounted for the average expected boost in performance for a given test. Several studies have instead proposed using a lack of PEs at the level of individuals to identify impairment and/or predict risk for AD and future cognitive decline [23]. In such studies, a lack of PEs is often defined as participants who show smaller (or no) improvements in performance at follow-up. As we noted, individuals exhibiting performance declines may still be benefitting from practice, and it is not always clear whether relative differences compared to group trajectories reflect lack of practice or simply greater decline. Hence, it will be important for future studies seeking to estimate individual-level PEs to develop robust methods that can disentangle practice effects from decline. Other studies employ multiple tests within short time frames (e.g., 1 week test-retest interval) and use lack of improved performance as indicators of risk for MCI or cognitive decline [49,50]. Although bearing a similar name, these short-term PE studies have a different purpose than our approach. These studies focus more directly on rate of learning with the goal of prognosis, or predicting subsequent decline. In contrast, our approach focuses on adjusting follow-up scores to obtain a more accurate assessment of current performance, which can enable earlier detection of impairment if PEs were sufficient to push an individual's scores above the threshold of impairment.

There are likely to be multiple mechanisms underlying practice effects, with varying contributions depending on the test. Participants may explicitly remember certain content from prior assessments (e.g., aspects of the story given during Logical Memory), which aids their performance. Yet on some tests such as Digit Span, it is highly unlikely participants are remembering specific test items (e.g., sequences of digits). In these scenarios, practice effects may be driven by greater familiarity with the *context.* This could include reduced test anxiety due to familiarity with the testing environment, procedural memory for tasks with a motor component, or having identified effective test-taking strategies [51,52]. These context-related factors may

explain why even individuals with severe episodic memory impairment can still benefit from practice [48].

We note some limitations of the current study. The VETSA is an all-male sample and primarily non-Hispanic White, limiting its generalizability to other samples. PEs have been shown to differ across certain participant characteristics, including general cognitive ability, age, diagnostic status, and presence of pathology [18,21,53], yet other studies find no differences across a number of demographic characteristics such as age, sex, or race/ethnicity [54]. It stands to reason that as the cognitive processes underpinning PE change, so will the resulting boosts in performance. Moreover, study characteristics such as sample age, testing interval, and the specific measures used to assess each domain may impact the magnitude of practice effects [22], and our results provide additional support for this being the case. For these reasons, it is ideal to estimate cohort-specific and test-specific practice effects. In the case of VETSA, recruiting a relatively small number of replacement participants (~10% of the sample) was sufficient to estimate and adjust for PEs with meaningful impacts on measures of cognition and rates of MCI. Our prior results strongly demonstrated that the additional cost of recruiting these participants is outweighed by the benefits, including the potential for substantial cost savings in clinical trials [16]. Importantly, we have also previously implemented an approach that studies without dedicated replacement participants can use to obtain “pseudo-replacements” [16]. These pseudo-replacements then allow for application of replacement-based methods such as those described by Ronnlund et al. [12] or in the current study.

In summary, our findings demonstrate that PEs arising from repeated cognitive testing may persist across multiple assessments and over periods as long as 20 years. Significant PEs were found even when observed scores declined over time, and these PEs simply obscured the true extent of decline in multiple cognitive domains. Adjusting for PEs had a substantial impact on rates of MCI, with up to a 29% increase in prevalence after adjustment. We have found that PE-adjustment using AR-participant approaches reduces rates of reversion from MCI to

unimpaired and results in greater concordance between MCI diagnoses and biomarker-positivity, providing evidence that the increased prevalence of MCI is not driven by false positive diagnoses. Instead, PE-adjustment enables earlier detection of progression to MCI. These results underscore the importance of estimating and adjusting for PEs in longitudinal studies of aging and clinical trials aimed at slowing cognitive decline.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Aging at the National Institutes of Health (grant numbers R01s AG076838, AG022381, AG050595, AG064955; and K01 AG063805). The content of this manuscript 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; the Institute for Survey Research, Temple University. The authors gratefully acknowledge the continued cooperation of the twins and the efforts of many staff members.

DISCLOSURES

AMD is a founder and holds equity in CorTechs Laboratories, Inc., and serves on its Scientific Advisory Board. He is a member of the Scientific Advisory Board of Human Longevity, Inc., and receives funding through research agreements with GE HealthCare and Medtronic. The terms of this arrangement have been reviewed and approved by the University of California San Diego in accordance with its conflict of interest policies. All other authors report no financial interests or potential conflicts of interest.

DATA AVAILABILITY

Instructions for data access requests are available on the VETSA website (https://psychiatry.ucsd.edu/research/programs-centers/vetsa/researchers.html). Access to data from military induction can be requested from the Vietnam Era Twin Registry (https://www.seattle.eric.research.va.gov/VETR/Investigator_Access.asp).

REFERENCES

[1] Schaie KW. The course of adult intellectual development. Am Psychol. 1994;49:304–13.
[2] Schaie KW. Intellectual development in adulthood: The Seattle longitudinal study: Cambridge University Press; 1996.
[3] Schaie KW, Strother CR. A cross-sequential study of age changes in cognitive behavior. Psychological bulletin. 1968;70:671–80.
[4] Nesselroade JR, Baltes PB. Longitudinal research in the study of behavior and development: Academic Press New York; 1979.
[5] Miller JB, Cummings J, Nance C, Ritter A. Neuroscience learning from longitudinal cohort studies of Alzheimer's disease: Lessons for disease-modifying drug programs and an introduction to the Center for Neurodegeneration and Translational Neuroscience. Alzheimers Dement (N Y). 2018;4:350–6.
[6] Galasko D, Corey-Bloom J, Thal LJ. Monitoring progression in Alzheimer's disease. J Am Geriatr Soc. 1991;39:932–41.
[7] McCaffrey RJ, Ortega A, Haase RF. Effects of repeated neuropsychological assessments. Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists. 1993;8:519–24.
[8] McCaffrey RJ, Ortega A, Orsillo SM, Nelles WB, Haase RF. Practice effects in repeated neuropsychological assessments. Clinical Neuropsychologist. 1992;6:32–42.
[9] Thorndike EL. Practice Effects in Intelligence Tests. J Exp Psychol. 1922;5:101–7.
[10] Rabbitt P, Diggle P, Holland F, McInnes L. Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. J Gerontol B Psychol Sci Soc Sci. 2004;59:P84–97.
[11] 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–P40.
[12] Ronnlund 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.
[13] Jacobs DM, Ard MC, Salmon DP, Galasko DR, Bondi MW, Edland SD. Potential implications of practice effects in Alzheimer's disease prevention trials. Alzheimers Dement (N Y). 2017;3:531–5.
[14] 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.
[15] Elman JA, Jak AJ, Panizzon MS, Tu XM, Chen T, Reynolds CA, et al. Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. Alzheimers Dement (Amst). 2018;10:372–81.
[16] Sanderson-Cimino M, Elman JA, Tu XM, Gross AL, Panizzon MS, Gustavson DE, et al. Cognitive practice effects delay diagnosis of MCI: Implications for clinical trials. Alzheimers Dement (N Y). 2022;8:e12228.
[17] Sanderson-Cimino M, Elman JA, Tu XM, Gross AL, Panizzon MS, Gustavson DE, et al. Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments. Front Aging Neurosci. 2022;14:847315.
[18] Mitrushina M, Satz P. Effect of repeated administration of a neuropsychological battery in the elderly. J Clin Psychol. 1991;47:790–801.
[19] Machulda MM, Pankratz VS, Christianson TJ, Ivnik RJ, Mielke MM, Roberts RO, et al. Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. Clin Neuropsychol. 2013;27:1247–64.
[20] Sanderson-Cimino M, Chen R, Tu XM, Elman JA, Jak AJ, Kremen WS. Misinterpreting

cognitive change over multiple timepoints: When practice effects meet age-related decline. Neuropsychology. 2023;37:568–81.
[21] Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology. 2010;24:563–72.
[22] 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.
[23] Jutten RJ, Grandoit E, Foldi NS, Sikkes SAM, Jones RN, Choi SE, et al. Lower practice effects as a marker of cognitive performance and dementia risk: A literature review. Alzheimers Dement (Amst). 2020;12:e12055.
[24] Schaie KW. Developmental influences on adult intelligence: The Seattle longitudinal study: Oxford University Press; 2005.
[25] Kremen WS, Franz CE, Lyons MJ. VETSA: the Vietnam Era Twin Study of Aging. Twin Res Hum Genet. 2013;16:399–402.
[26] Kremen WS, Franz CE, Lyons MJ. Current Status of the Vietnam Era Twin Study of Aging (VETSA). Twin Res Hum Genet. 2019;22:783–7.
[27] Kremen WS, Thompson-Brenner H, Leung YM, Grant MD, Franz CE, Eisen SA, et al. Genes, environment, and time: the Vietnam Era Twin Study of Aging (VETSA). Twin Res Hum Genet. 2006;9:1009–22.
[28] 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.
[29] Goldberg J, Curran B, Vitek ME, Henderson WG, Boyko EJ. The Vietnam Era Twin Registry. Twin research : the official journal of the International Society for Twin Studies. 2002;5:476–81.
[30] 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.
[31] Tsuang MT, Bar JL, Harley RM, Lyons MJ. The Harvard Twin Study of Substance Abuse: what we have learned. Harv Rev Psychiatry. 2001;9:267–79.
[32] Schoenborn CA, Heyman KM. Health characteristics of adults aged 55 years and over: United States, 2004-2007. Natl Health Stat Report. 2009;16:1–31.
[33] Bayroff A, Anderson AA. Development of the Armed Forces Qualification Tests 7 and 8 (Technical Research Report 1122). In: Institute. UAR, editor. Alexandria, VA1963.
[34] 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.
[35] Tang W, He H, Tu XM. Applied categorical and count data analysis: Chapman and Hall/CRC; 2023.
[36] Gustavson DE, Elman JA, Sanderson-Cimino M, Franz CE, Panizzon MS, Jak AJ, et al. Extensive memory testing improves prediction of progression to MCI in late middle age. Alzheimers Dement (Amst). 2020;12:e12004.
[37] Gustavson DE, Panizzon MS, Elman JA, Franz CE, Beck A, Reynolds CA, et al. Genetic and Environmental Influences on Verbal Fluency in Middle Age: A Longitudinal Twin Study. Behav Genet. 2018;48:361–73.
[38] Gustavson DE, Panizzon MS, Franz CE, Friedman NP, Reynolds CA, Jacobson KC, et al. Genetic and environmental architecture of executive functions in midlife. Neuropsychology. 2018;32:18–30.
[39] Sanderson-Cimino M, Panizzon MS, Elman JA, Gustavson DE, Franz CE, Reynolds CA, et al. Genetic and environmental architecture of processing speed across midlife. Neuropsychology. 2019;33:862–71.
[40] 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. Journal of Alzheimer's disease : JAD.

2014;42:275–89.
[41] 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.
[42] 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.
[43] Machulda MM, Hagen CE, Wiste HJ, Mielke MM, Knopman DS, Roberts RO, et al. Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. Clin Neuropsychol. 2017;31:99–117.
[44] 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. Journal of the International Neuropsychological Society : JINS. 2003;9:419–28.
[45] 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.
[46] Ferrer E, Salthouse TA, Stewart WF, Schwartz BS. Modeling age and retest processes in longitudinal studies of cognitive abilities. Psychol Aging. 2004;19:243–59.
[47] Tort-Merino A, Perez-Millan A, Falgas N, Borrego-Ecija S, Esteller D, Bosch B, et al. Decreased practice effects in cognitively unimpaired amyloid betapositive individuals: a multicenter, longitudinal, cohort study. Alzheimers Dement. 2025;21:e70016.
[48] Gross AL, Chu N, Anderson L, Glymour MM, Jones RN, Coalition Against Major D. Do people with Alzheimer's disease improve with repeated testing? Unpacking the role of content and context in retest effects. Age Ageing. 2018;47:866–71.
[49] Duff K, Atkinson TJ, Suhrie KR, Dalley BC, Schaefer SY, Hammers DB. Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. J Clin Exp Neuropsychol. 2017;39:396–407.
[50] 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.
[51] Anastasi A, Urbina S. Psychology testing. New Jersey: Prentice Hall; 1997.
[52] Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology. 2008;22:800–11.
[53] Reeve CL, Lam H. The Relation Between Practice Effects, Test-Taker Characteristics and Degree ofg-Saturation. International Journal of Testing. 2007;7:225–42.
[54] Gross AL, Benitez A, Shih R, Bangen KJ, Glymour MM, Sachs B, et al. Predictors of Retest Effects in a Longitudinal Study of Cognitive Aging in a Diverse Community-Based Sample. Journal of the International Neuropsychological Society : JINS. 2015;21:506–18.
[55] Golden CJ. Stroop color and word test: A manual for clinical and experimental uses. Chicago, IL: Stoelting; 1978.
[56] Daneman M, Carpenter PA. Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior. 1980;19:450–66.
[57] Ekstrom RB, French JW, Harman HH, Dermen D. Manual for kit of factor-referenced cognitive tests. Princeton, NJ: Educational testing service; 1976.
[58] Thurstone LL. A factorial study of perception. Chicago, IL: University of Chicago Press; 1944.
[59] Delis DC, Kramer JH, Kaplan E, Ober BA. California Verbal Learning Test. Second Edition. San Antonio, TX: Psychological Corporation; 2000.
[60] Wechsler D. Wechsler Adult Intelligence Scale-III (WAIS-III) manual. San Antonio, TX: The Psychological Corporation.; 1997.
[61] Delis DC, Kaplan E, Kramer JH. Delis-Kaplan executive function system: Technical manual. San Antonio, TX: Psychological Corporation; 2001.

[62] Wechsler D. Manual for the Wechsler abbreviated intelligence scale (WASI). San Antonio, TX: The Psychological Corporation. 1999.

TABLES & FIGURES

Table 1. Demographic characteristics of the sample.

N Total	1608
Age (years); mean (SD)	
Wave 1	55.94 (2.43)
Wave 2	61.72 (2.45)
Wave 3	67.56 (2.53)
Wave 4	73.09 (2.09)
Education (years); mean (SD)	13.85 (09)
Age 20 AFQT (percentile); mean (SD)	59.95 (23.11)
Race; n (%)	
American Indian or Alaskan Native	5 (0.3%)
Native Hawaiian or Pacific Islander	2 (0.1%)
Black or African-American	88 (5.5%)
White	1492 (92.8%)
More than one race	21 (1.3%)
Decline to answer	2 (0.1%)
Ethnicity; n (%)	
Hispanic	44 (2.7%)
Non-Hispanic	1564 (97.3%)
Testing interval (years); mean (SD)	
Wave 1 to 2	5.69 (0.69)
Wave 2 to 3	5.91 (0.76)
Wave 3 to 4	5.63 (0.60)

Note: AFQT = Armed Forces Qualifying Test

Table 2. Neuropsychological tests and scores included in practice effects estimation.

Cognitive Domain	Measure
General Cognitive Ability	AFQT – Total Score
Episodic Memory	CVLT-II Total of Trials 1-5 CVLT-II Short Delay Free Recall CVLT-II Long Delay Free Recall WMS-III Logical Memory - Immediate Recall Story Units Total Score WMS-III Logical Memory - Delayed Recall Story Units Total Score
Executive function	D-KEFS Trails 4 time (adjusted for Trails 2 & 3) D-KEFS Category switching accuracy (adjusted for verbal fluency) Stroop: Interference Score (adjusted for color & word) [55]
Verbal Fluency	D-KEFS Letter Fluency FAS total correct D-KEFS Category Fluency Animal/Boys total correct
Processing Speed	D-KEFS Trails 2 time D-KEFS Trails 3 time Stroop: Raw Word Score [55] Stroop: Raw Color Score [55] Simple Reaction Time Choice Reaction Time
Short-term/Working Memory	WMS-III Digit Span: Forward + Backward Score WMS-III Spatial Span: Total Trials Passed Forward + Backward WMS-III Letter-Number Sequencing: Total Score for Trials Passed Reading Span: Total Score Ascending [56]
Visuospatial	Mental Rotation: Total Correct Part 1 [57] Gottschaldt Hidden Figures Total Correct All Parts [58] AFQT – Box Folding Subtest WMS-III Visual Reproduction: Copy Total Score
Visual Memory	WMS-III Visual Reproduction: Immediate Recall Total Score WMS-III Visual Reproduction: Delayed Recall Total Score
Abstract Reasoning	WASI Matrix Reasoning Raw Score

Note: AFQT = Armed Forces Qualifying Test [33]; CVLT = California Verbal Learning Test-II [59]; WMS-III = Wechsler Memory Scale-III [60]; D-KEFS = Delis-Kaplan Executive Function System [61]; WASI = Wechsler Abbreviated Scale of Intelligence [62]

Figure 1. Summary of assessments and age distributions by wave. A) Upset plot describing the patterns of baseline and follow-up assessments completed. Individuals with baseline assessments at waves 2 and 3 were recruited as attrition replacements. B) Density plots of age distributions at each wave. Attrition replacement (AR) participants were recruited at waves 2 and 3 and were age-matched to the on-going longitudinal sample at the corresponding timepoint.

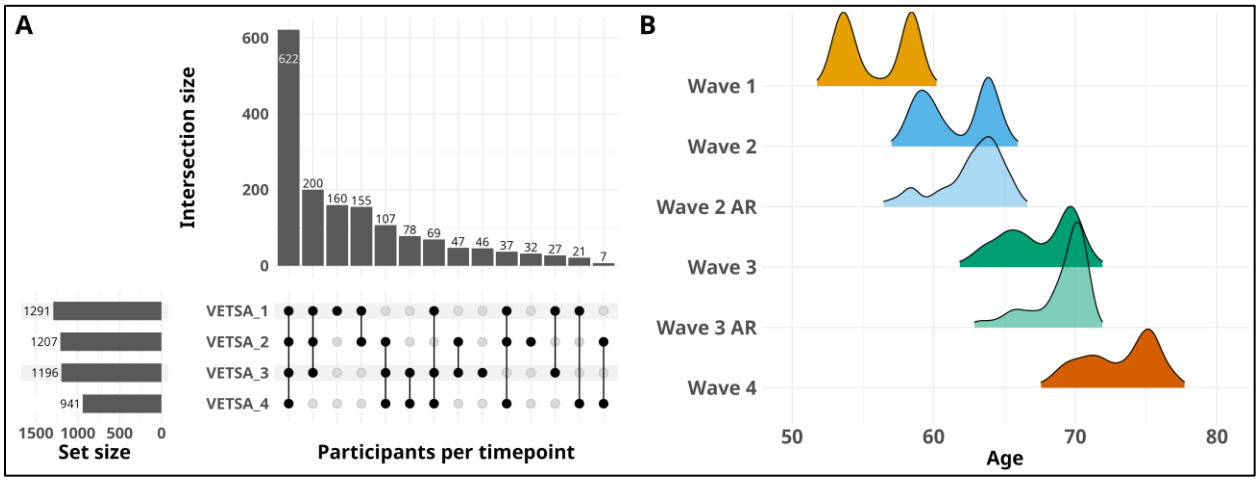

Figure 2. Practice effect estimates across follow-up assessments. The forest plot presents practice effect estimates and 95% confidence intervals at each follow-up assessment (i.e., waves 2, 3 and 4). The presented estimates are for individuals that participated in all four study waves such that estimates reflect practice effects resulting from completing all prior assessments. Dots and whiskers for each measure are colored according to the cognitive domain that they assess. All items were coded such that higher values reflect better performance. See Table 2 for full names of measures in each domain. GCA = general cognitive ability; AFQT = Armed Forces Qualifying Test

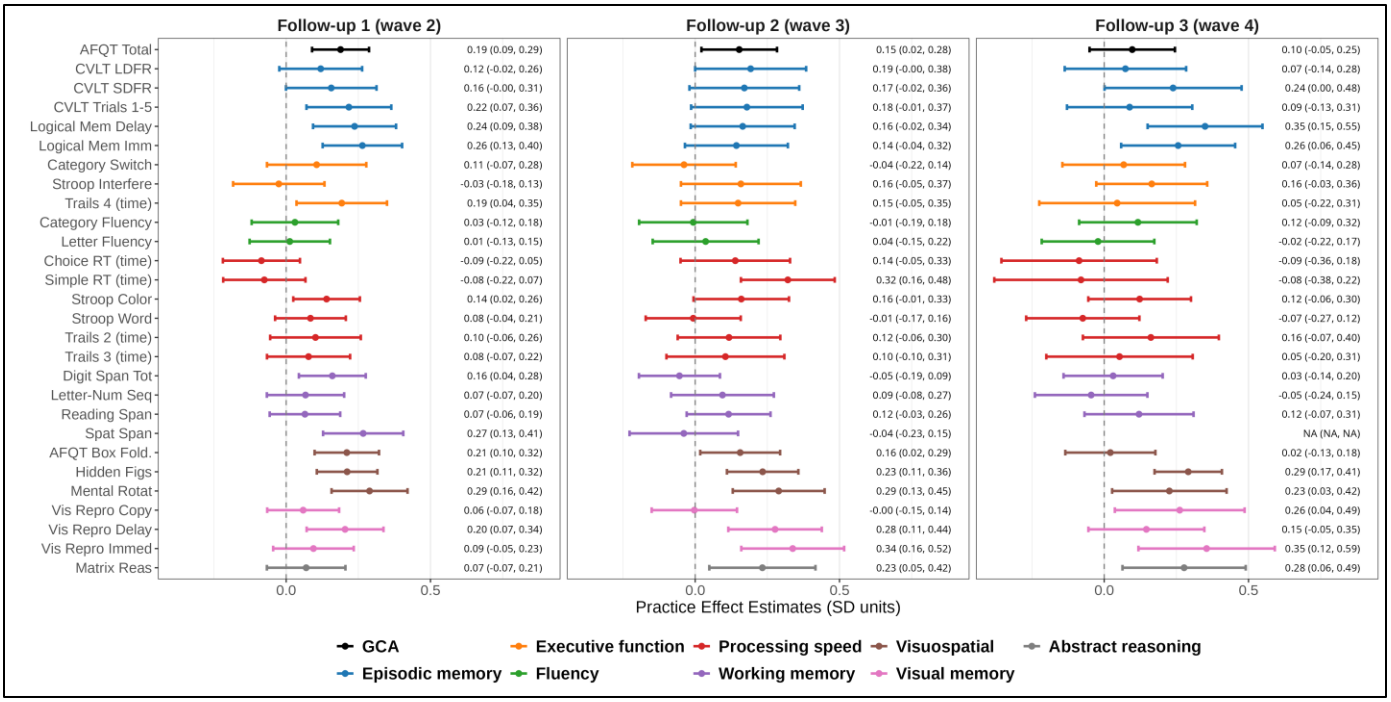

Figure 3. Plots of cognitive factor score trajectories. Means and 95% confidence intervals for cognitive factor score composites calculated from unadjusted (triangles and dashed lines) versus practice effect-adjusted (dots and solid lines) scores. Confidence intervals were calculated using within-subject standard errors. Scores were standardized using the sample means and standard deviations at wave 1.

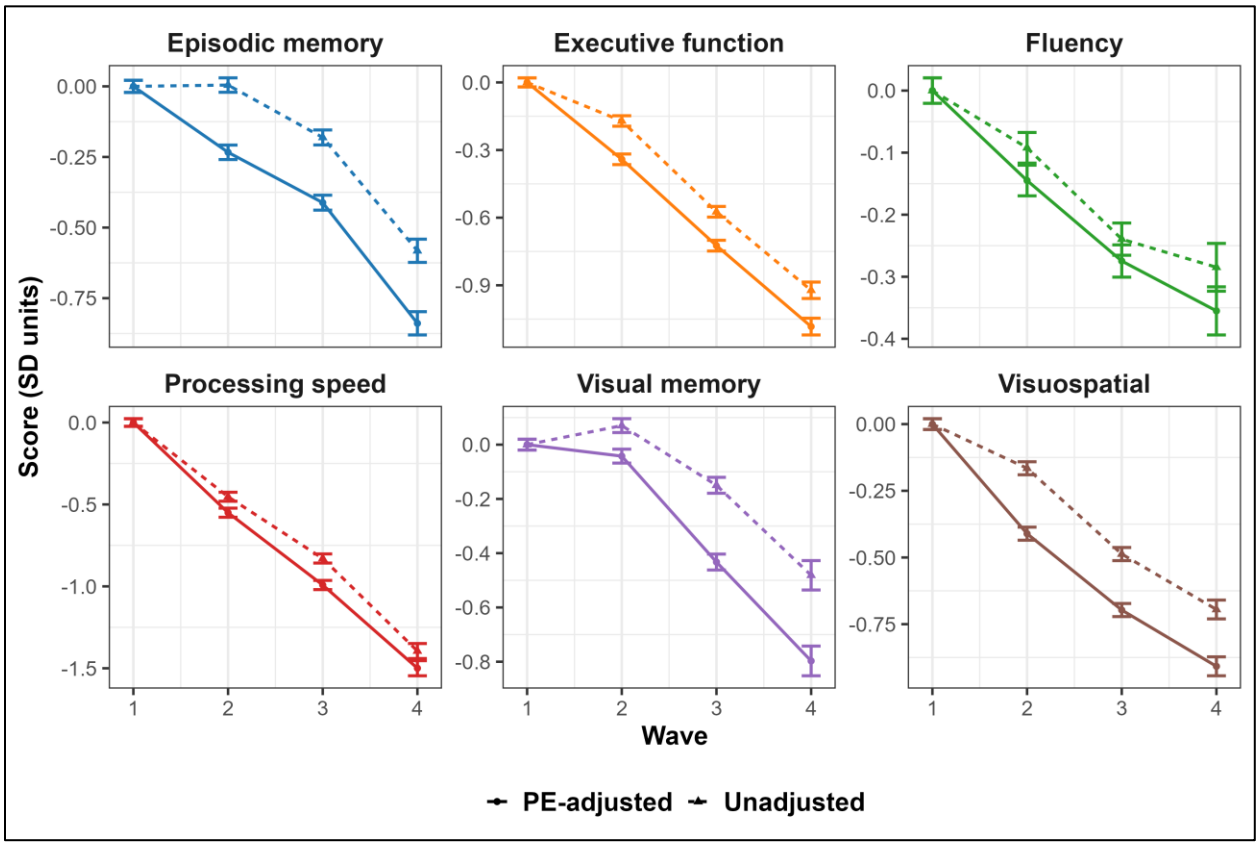
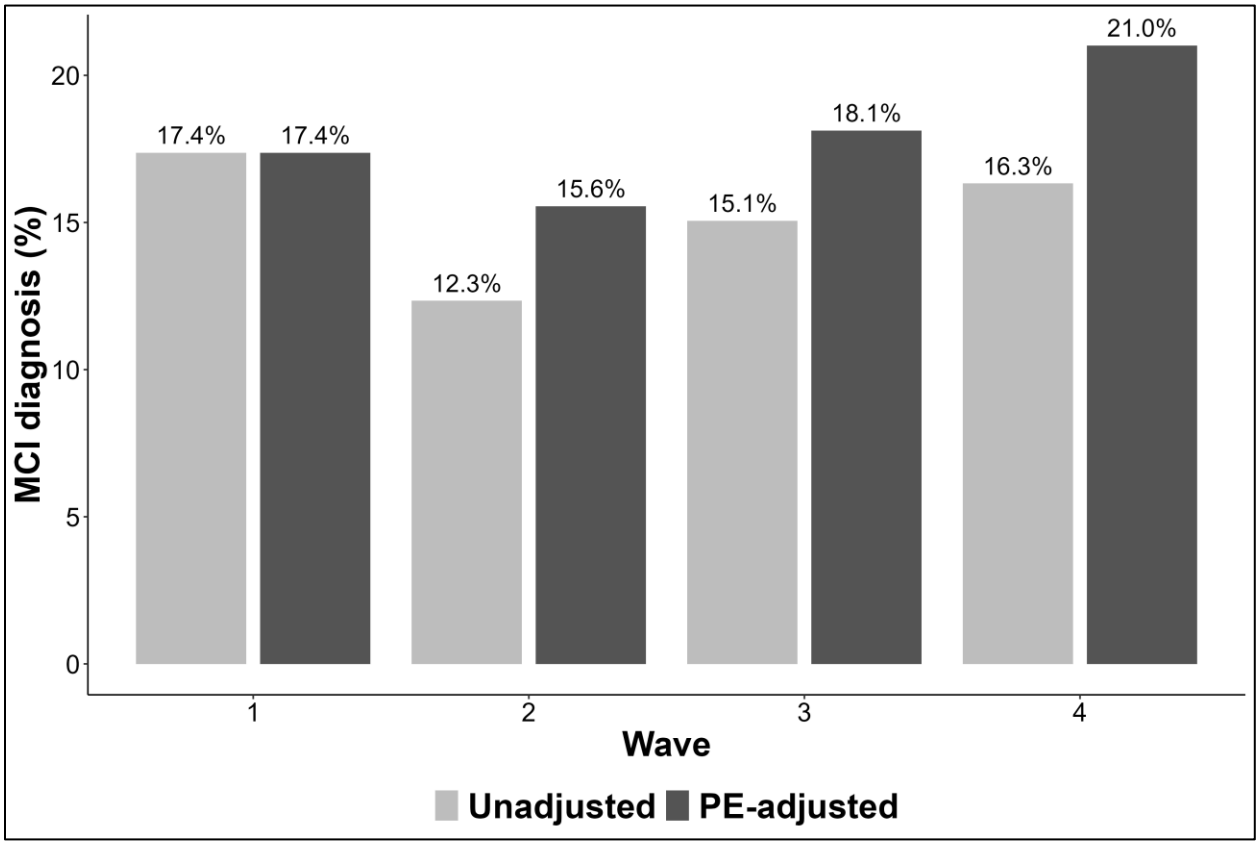
Figure 4. Rates of mild cognitive impairment. Bar plots present prevalence of mild cognitive impairment (both amnestic and non-amnestic) diagnoses based on unadjusted (light grey) and practice effect-adjusted (dark gray) test scores.