AMERICAN PSYCHOLOGICAL ASSOCIATION

Neuropsychology

ISSN: 0894-4105

2023, Vol. 37, No. 5, 568–581
https://doi.org/10.1037/neu0000903

Misinterpreting Cognitive Change Over Multiple Timepoints: When Practice Effects Meet Age-Related Decline

Mark Sanderson-Cimino[1, 2], Ruohui Chen[3], Xin M. Tu[4, 5, 6], Jeremy A. Elman[2, 4], Amy J. Jak[2, 7], and William S. Kremen[2, 4]

[1] Memory and Aging Center, University of California San Francisco
[2] Center for Behavior Genetics of Aging, University of California San Diego
[3] Division of Biostatistics and Bioinformatics, Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego
[4] School of Medicine, University of California San Diego
[5] Family Medicine and Public Health, University of California San Diego
[6] Sam and Rose Stein Institute for Research on Aging, University of California San Diego
[7] Center of Excellence for Stress and Mental Health, Veterans Affairs San Diego Healthcare System, San Diego, California, United States

Objective: Practice effects (PE) on cognitive testing have been shown to delay detection of impairment and impede our ability to assess change. When decline over time is expected, as with older adults or progressive diseases, failure to adequately address PEs may lead to inaccurate conclusions because PEs artificially boost scores while pathology- or age-related decline reduces scores. Unlike most methods, a participant-replacement approach can separate pathology- or age-related decline from PEs; however, this approach has only been used across two timepoints. More than two timepoints make it possible to determine if PEs level out after the first follow-up, but it is analytically challenging because individuals may not be assessed at every timepoint. ***Method:*** We examined 1,190 older adults who were cognitively unimpaired ($n = 809$) or had mild cognitive impairment (MCI; $n = 381$). Participants completed six neuropsychological measures at three timepoints (baseline, 12-month, 24-month). We implemented a participant-replacement method using generalized estimating equations in comparisons of matched returnees and replacements to calculate PEs. ***Results:*** Without accounting for PEs, cognitive function appeared to improve or stay the same. However, with the participant-replacement method, we observed significant PEs within both groups at all timepoints. PEs did not uniformly decrease across time; some—specifically on episodic memory measures—continued to increase beyond the first follow-up. ***Conclusion:*** A replacement method of PE adjustment revealed significant PEs across two follow-ups. As expected in these older adults, accounting for PEs revealed cognitive decline. This, in turn, means earlier detection of cognitive deficits, including progression to MCI, and more accurate characterization of longitudinal change.

Key Points

Question: How do practice effects on cognitive tests change over time, and how do they impact understanding of cognitive aging trends? ***Findings:*** Practice effects may not level out after the first follow-up visit and their inclusion in analyses leads to change that is more in line with expected cognitive aging trends. ***Importance:*** Failure to properly account for practice effects leads to inaccurate conclusions about cognitive change. ***Next Steps:*** Future studies should investigate practice effects over longer time periods and with samples that are more representative of the general population.

Keywords: practice effects, cognitive aging, longitudinal change

Supplemental materials: https://doi.org/10.1037/neu0000903.supp

This article was published Online First April 20, 2023.

Mark Sanderson-Cimino https://orcid.org/0000-0002-5872-3186

Mark Sanderson-Cimino and Ruohui Chen are both first authors.

Mark Sanderson-Cimino and William S. Kremen received funding from Grants F31 AG064834, R01 AG050595, R01 AG076838, R01 AG064955, R01 AG060470, respectively, from the National Institute on Aging.

Mark Sanderson-Cimino played a lead role in conceptualization, methodology and writing. Ruohui Chen played lead role in methodology. Xin M. Tu played supporting role in formal analysis and methodology. Jeremy A. Elman played supporting role in conceptualization, methodology and writing–review and editing. Amy J. Jak played supporting role in writing–review and editing. William S. Kremen played supporting role in conceptualization, funding acquisition, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Mark Sanderson-Cimino, Memory and Aging Center, University of California San Francisco, Box 1207, 675 Nelson Rising Lane, Suite 190, San Francisco, CA 94143, United States. Email: mark.sandersoncimino@ucsf.edu

Some cognitive change over time is expected as adults age, particularly in those over the age of 65 (Finkel et al., 2003; Salthouse, 2010, 2019). Those with mild deficits in cognitive domains, beyond what would be expected for aging, may be diagnosed with mild cognitive impairment (MCI), which is seen as a prodromal stage of dementia (Albert et al., 2011; Eppig et al., 2017; Manly et al., 2008; Thomas et al., 2020). If an individual progresses to greater cognitive impairment accompanied by substantial declines in their daily functioning, they may meet criteria for major neurocognitive disorder (i.e., dementia; Albert et al., 2011; Eppig et al., 2017; Manly et al., 2008; Thomas et al., 2020). As normal and abnormal aging are inherently longitudinal processes, repeated assessments are essential for diagnoses and mapping change over time.

Despite the need for and use of repeated testing, cognitive diagnoses are almost always made with respect to the *most recent* assessment, without considering how prior testing may have influenced results (Calamia et al., 2012; Goldberg et al., 2015; Heilbronner et al., 2010). Repeated assessments are subject to practice effects (PEs) that impair our ability to detect change. PEs can be defined as improvements in performance due to familiarity with testing rather than any actual alteration of true ability; put simply, someone taking a cognitive test for the second time often does better than if they were taking it for the first time (Calamia et al., 2012; Heilbronner et al., 2010; Salthouse, 2019; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). PEs are sometimes separated into content (i.e., knowledge of specific stimuli) and context (i.e., improved familiarity with testing, reduced anxiety) effects, although this delineation is somewhat heuristic (Gross et al., 2018; Heilbronner et al., 2010).

PEs are often ignored or minimally addressed in both research and clinical settings (Calamia et al., 2012; Goldberg et al., 2015; Heilbronner et al., 2010; Machulda et al., 2017; Mathews et al., 2014; Salthouse, 2019; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). This is somewhat alarming as PEs are pervasive, occurring across all cognitive domains, and long-lasting, with studies noting PEs after up to 7 years postbaseline (Elman et al., 2018; Goldberg et al., 2015; Gross et al., 2015, 2017; Rönnlund et al., 2005). Moreover, they have been found in individuals who at baseline are cognitively unimpaired (CU), diagnosed with MCI, and even in those with mild Alzheimer's disease (AD; Elman et al., 2018; Goldberg et al., 2015; Gross et al., 2015, 2017; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). It has been suggested that addressing PEs might lead to an earlier diagnosis of impairment (Goldberg et al., 2015; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022), but to our knowledge, only one method of estimating PEs has the possibility of earlier diagnosis essentially built into it—the participant-replacement method (Elman et al., 2018; Rönnlund et al., 2005). Using this method, it has been shown that adjusting for PEs leads to an earlier detection of MCI, improves stability of MCI diagnoses, and strengthens our ability to predict conversion to dementia (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). Failure to adjust for PEs also decreases statistical power and may have a substantial impact on the financial, staff, and patient burden of clinical drug trials that investigate conversation to MCI or dementia (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022).

Most studies only label increases in scores as PEs, meaning that PEs are only identified if an individual has a higher score at follow-up than at baseline (Calamia et al., 2012; Goldberg et al., 2015). This definition is problematic within populations expected to experience cognitive decline such as older adults or those on the AD trajectory as PEs improve scores over time while neurodegeneration worsens scores, particularly over longer retest intervals. Across short retest intervals (e.g., 1 week from baseline) PEs in AD and other neurodegenerative populations are easier to identify. In that case there is often a clear improvement in scores at retesting as it would be very unlikely for an individual to experience significant neurodegeneration over 1 week that would be greater than their PE (Duff, 2014; Duff et al., 2011, 2014). Moreover, studies with short retest intervals are designed to maximize PEs, with the goal of creating a measure of consolidation that can be used to predict who is at risk of future decline (Duff, 2014; Saloner et al., 2018). This paradigm can provide extremely useful information, but it does not address the issue that PEs can interfere with detection of true change, particularly in studies of older adults. Short term PEs also can not affect when progression to MCI is detected, which to our knowledge, is only possible with the replacement method (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). Finally, we note that PE research appears to be very susceptible to the jingle fallacy, the erroneous assumption that different things are the same because they have the same or equivalent label (Thorndike, 1913). Hence, "which approach is best?" is a common, but not an appropriate, question. Although the label "practice effects" is the same for these different approaches, it is important to note that each addresses very different issues (Kremen et al., 2022).

Longer retest intervals (e.g., 6 months or greater) might involve PEs that are equal to or less than the true change over time. For example, if an individual truly declines two points on a memory measure between annual assessments, but experiences a PE of three points, then they will appear to improve by 1 point as they age. True improvement is unlikely in those with a neurodegenerative disease, and improved test performance is often considered a PE. However, if that same individual instead experiences a PE of only 1 point, they will appear to decline by 1 point. The key idea here is that there is still a PE despite the fact that performance declines, because the true 2-point decline is masked, appearing as only a 1-point change. This latter situation can occur but is typically missed as most approaches to PEs do not allow for simultaneous modeling of PEs and age-related decline (Calamia et al., 2012; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022). Specifically, standardized regression-based approaches and reliable change indexes (RCIs) that incorporate PEs define no or nonsignificant change as a lack of PE (Chelune et al., 1993). This is not optimal for populations where decline over time is expected (e.g., older adults; Duff, 2012).

Studies with multiple timepoints typically conclude that the magnitude of a PE levels out over time, particular after the first retesting (Calamia et al., 2012; Goldberg et al., 2015). This idea is based on the observation that scores tend to increase less over multiple follow-up visits. While this observation may be partly due to diminishing PEs, it may also be that the effect of age-related

decline over multiple follow-up years outpaces the PEs, reducing the observable increase in scores. To our knowledge, this hypothesis has been raised, but never formally tested.

A recent review of PEs in neurological disorders noted that the replacement method has not been applied across multiple retest visits, which the review labeled as a key limitation of the approach (Holm et al., 2022). Application of the method to multiple timepoints would address this concern while also commenting on the evolution of PEs alongside expected age-related decline. In the present analyses, we used a modified version of the participant-replacement method of PE adjustment (see Method section; Elman et al., 2018; Rönnlund et al., 2005; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022) and generalized estimating equations (GEE) to examine PEs at more than two visits: baseline, 12-month follow-up, and 24-month follow-up. Participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were diagnosed as CU or MCI at baseline. Models were completed separately for these two subsamples across six neuropsychological measures. Models were completed first without adjusting for PEs and second after adjusting for PEs. We hypothesized that (a) PEs will be present at both the 12-month follow-up and the 24-month follow-up; (b) PEs will increase across time; (c) the PE-adjusted models will find both significant age-related decline and significant PEs; (d) PE-unadjusted models will provide inaccurate estimates of change as compared to the PE-adjusted models.

Method

Transparency and Openness

We report all data exclusions, manipulations, and measures below. Data was provided by ADNI (see below) and is available at https://adni.loni.usc.edu/. This study design and its analysis were not preregistered. Analyses were completed in R (Team, 2021).

Participants

All participants were enrolled in the ADNI (https://adni.loni.usc.edu) which was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information and access to data, see https://www.adni-info.org. Participants from the ADNI-1, ADNI-GO, and ADNI-2 cohorts were included.

There were 1,190 participants with baseline data who did not meet ADNI's criteria for dementia at study entry. A diagnosis of MCI (amnestic, dysexecutive, or language-impaired) was made using the Jak-Bondi approach (Jak et al., 2009). Participants were classified as single domain MCI if their scores on two tests within the same cognitive domain were both greater than 1 *SD* below normative means. They were diagnosed as multidomain MCI if they met the impairment criteria in more than one cognitive domain (Jak et al., 2009). All participants completed at least one neuropsychological measure at baseline.

We investigated PEs across three visits: baseline, 12-month follow-up, and 24-month follow-up. There were 858 participants with cognitive data at all visits. There were 258 participants with data at baseline and the 12-month follow-up, but no 24-month data. There were 74 participants with baseline and the 24-month follow-up, but no 12-month data.

Measures

Participants completed up to six neuropsychological measures at each visit. Episodic memory tasks included the Wechsler Memory Scale–Revised, Logical Memory Story A delayed recall (Chelune et al., 1990), and the Rey Auditory Verbal Learning Task (AVLT) delayed recall (Schmidt, 1996). Language tasks included the Boston Naming Test (BNT; Kaplan et al., 2001) and Category (Animal) Fluency (Petersen et al., 2010). Attention–executive function tasks were Trails A and Trails B (Lezak et al., 2004). The American National Adult Reading Test (ANART) provided an estimate of premorbid IQ (Taylor et al., 1996). Although ADNI included alternate forms at some visits, all participants in the present study completed the same version of the tests at each visit. Of note, scores on Trails A and B were reversed so that more positive scores indicate better performance on all tests.

Statistical Analyses

For use in studies such as ADNI that did not originally recruit replacement participants, we adapted the method by creating the equivalent of replacement participants (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). We identified a subsample of baseline participants who are demographically matched to the returnees at follow-up and labeled as "pseudoreplacements." Propensity scores were used to ensure that the pseudoreplacements and returnees were similar, with respect to age and other demographic characteristics. A comparison of the pseudoreplacement scores with the returnee scores yields a PE because the only significant difference between these groups is that the returnees have taken the test before and the pseudoreplacements have not. Importantly, once the groups have been matched, pseudoreplacements are functionally the same as replacement participants who are recruited for that purpose, as part of a study's original design. As such, we will use the term "replacements" in lieu of "pseudoreplacements" moving forward.

The replacement method is essentially a longitudinal approach that uses propensity score matching with a cross-sectional subsample to reduce cohort effects. This reduction allows for a more accurate quantification of expected or normative longitudinal change (Salthouse, 2019) across a given time period (baseline to follow-up assessment). Other methods of PE adjustment also incorporate a comparison group (e.g., RCIs, standardized regression), and when this group is not well matched to the target sample the PE estimates are inaccurate and may be unusable (Calamia et al., 2012; Duff & Hammers, 2022).

The replacement method has been used for PEs at a single retest visit (Elman et al., 2018; Rönnlund et al., 2005; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022), but it has not been used for multiple retest visits. A complication here is that some returnees attend all follow-up visits while others only attend some. As PEs may change based on the number of test exposures (Calamia et al., 2012; Stricker et al., 2020), it is

essential to also match on the number of prior visits. As such, it was necessary to identify multiple samples of replacement groups to match to the separate returnee groups.

To estimate PEs at the 12-month follow-up, we subsampled 30% of the returnees ($n_r = 257$). The remaining 70% ($n = 933$) include dropouts and returnees who did not attend all visits: 74 participants with only baseline data; 258 participants with only two visits; and 601 returnees that were not selected. This group of 933 served as a pool for potential replacements to calculate PEs at the 12-month follow-up. Propensity scores were used to select replacements from this pool (70% of sample; $n = 933$) that matched the randomly selected returnees (30% of the sample; $n = 257$) on age, sex, and the ANART. We define this group of replacements as Group A ($n = 257$); at their baseline, they had similar demographics to the subsampled returnees ($n_r = 257$) at their 12-month follow-up. To estimate the PEs at the 24-month follow-up, we followed a similar strategy. We retained the same returnee subsample (i.e., the randomly selected 30% of overall sample; $n_r = 257$). We returned to the replacement pool (70% of sample; $n = 933$) and matched the 257 returnees to 257 individuals with a baseline and 12-month follow-up visit (i.e., replacements). These 257 replacements were matched to the 257 returnees on age, sex, and the ANART using propensity scores. We define this group of replacements ($n = 257$) as Group B; they had completed tests at the 12-month follow-up and have similar demographics to the subsampled 30% of returnees at the returnees' 24-month follow-up. Importantly, this means that the returnees' age at their 24-month visit (i.e., 3rd test visit; $n = 257$) was matched to the age of the Group B replacements at their 12-month visit (i.e., 2nd test visit; $n = 257$). As a result, a comparison of returnee follow-up data at the 24-month follow-up to Group B's scores calculated the additional PEs from Visit 2 to Visit 3. Of note, participants were separated based on baseline cognitive status (MCI vs. CU), meaning that replacements had the same baseline diagnosis as their matched returnees. Figure 1 provides a graphical depiction of the participant matching procedure.

After creating the matched replacement groups' data, we used GEEs to estimate the time effects and PEs. The GEEs used the identity link function and the working independence correlation structure (Tang et al., 2012). We also performed sensitivity analyses by using different correlation structures, such as "exchangeable," "autoregressive," and "unstructured." As results were very similar, we report findings based on the working independence correlation structure. Although both GEE and linear mixed-effects models are applicable and able to provide population-average effects of interest, we opted for GEE because of its robust and optimal inference without imposing any mathematical model on the distribution of the response (Tsiatis, 2006).

Let "t" denote the timepoint at which participants take the test; "$t = 1$" represents the baseline, "$t = 2$" represents the 12-month follow-up, and "$t = 3$" represents the 24-month follow-up. "G" denotes group, and "$G = R$" represents returnees for the reference group; these have data at both follow-ups. Y_{it} denotes participant i's cognitive scores at time t, and X_{it} denotes a vector of covariates of participant i at time t. Therefore:

$$E[X_{it}] = \beta_0 + \beta_x X_{it} + \beta_1 I(t = 2) + \beta_2 I(t = 3) + \beta_3 I(t = 2)(G = A) + \beta_4 I(t = 2)(G = B), \quad (1)$$

where β_1 is the time effect at 12 months, β_2 is the time effect at 24 months, β_3 is the mean of PEs at the 12-month follow-up, and β_4 is the mean of PEs at the 24-month follow-up. We used the Wald test to determine the p values for the corresponding parameters. Therefore, the PE-unadjusted model ultimately included parameters for age, sex, education, the 12-month follow-up, and the 24-month follow-up. The PE-adjusted model included the same parameters as well as a PE at the 12-month visit and at the 24-month visit. Significance levels were set at $p < .05$.

Results

Sample Characterization

Table 1 provides a description of the CU and MCI groups as well as raw scores for each cognitive test. Within the group that was CU at baseline ($n = 809$), 735 (91%) returned for a 12-month visit and 691 (85%) returned at the time of the 24-month visit. Within the group that was MCI at baseline ($n = 381$), 381 (100%) returned for their 12-month visit and 241 (63%) returned at the time of their 24-month visit.

PEs Within the CU Group

At the 12-month follow-up visit, there were significant PEs across all measures. There were also significant PEs at the 24-month follow-up visit for five of the six measures. The PEs in raw score units are fully presented in Table 2. The significant PE estimates did not uniformly change over time. PEs on the Trails A test reduced over time by approximately 11% ([3.8–3.40]/3.81). Similarly, Trails B and the AVLT also declined over time (Trails B: −26%; AVLT −5%). However, PEs increased over time on Logical Memory (+39%) and the BNT (+41%).

PEs Within the MCI Group

At the 12-month follow-up visit, there were significant PEs across four of the six measures: Logical Memory, AVLT, Trails A, and Trails B. There were also significant PEs at the 24-month follow-up visit for two of the six measures: Logical Memory and Trails B. Regarding PEs that were significant at both timepoints, the Logical Memory PE was much larger at the 24-month follow-up as compared to the 12-month follow-up (+133%). The Trails B PE was reduced at the 24-month visit (−13%). The PEs in raw score units are fully presented in Table 3.

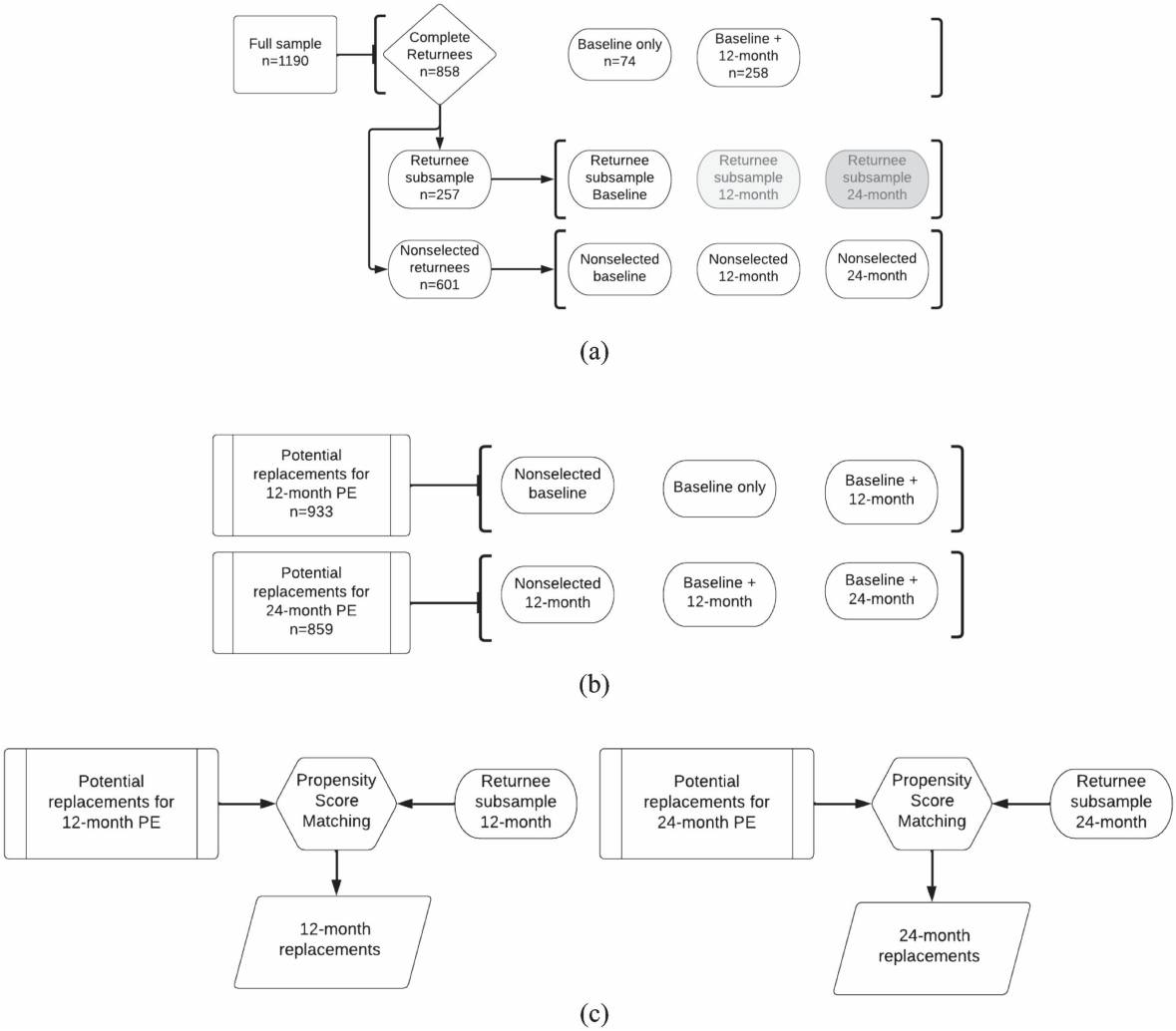
Cognitive Trajectories With and Without PE Adjustment in the CU Group

Within the CU group, the pattern of change-over-time was significantly different in the PE-adjusted and PE-unadjusted groups. Figure 2 displays the trajectories for PE-adjusted and PE-unadjusted scores for all cognitive measures among participants who were CU at baseline.

Logical Memory

Points on Logical Memory indicate correctly recalled story units. In the PE-unadjusted model, scores significantly improved as participants aged. They scored about 1.7 points higher at the 12-month follow-up and 1.1 points higher at the 24-month

Figure 1
Sample Composition and Replacement Matching



Note. (a) Sample composition: The full sample consisted of 1,190 participants. This included 858 participants with data at all three visits (complete returnees), 74 participants with only baseline data (baseline only), and 258 participants with data at only the first two visits (baseline + 12-month). A random subsample (30%) of complete returnees was selected (returnee subsample), leaving 601 participants (nonselected returnees). The returnee subsample and the nonselected returnee group all had data at all three visits. (b) Pseudoreplacement selection: There were 933 participants with baseline data that were not included in the returnee subsample group. The baseline visit for these participants was included in the potential replacement pool for the 12-month practice effect (PE). There were 859 participants with a baseline and 12-month visit that were not included in the returnee subsample group. The 12-month visit for these participants was included in the potential replacement pool for the 24-month PE. (c) Participant matching: The potential replacements for the 12-month PE were matched on age, sex, and ANART scores to the returnee subsample at the returnee's 12-month visit. This meant that the replacements were the same age as those in the returnee sample when the returnee sample completed the tests for the second time. Similarly, the potential replacements for the 24-month PE were matched to the returnee subsample at the returnee 24-month visit. One-to-one matching was used, as was a second constraint ensuring that groups were statistically similar after completion ($ps > .8$). As a result, the 12-month replacements ($n = 257$) were demographically identical to the returnees at their 12-month visit, and the 24-month replacements ($n = 257$) were demographically identical to the returnees at their 24-month visit. ANART = American National Adult Reading Test.

follow-up. In contrast, when PE estimates were included in the model, scores significantly worsened across time (−.54 points and −2.0 points, respectively). The PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 12-month follow-up ([1.4, 2.1] vs. [−.84, −.24]) and the 24-month follow-up ([.83, 1.4] vs. [−2.3, −1.7]).

AVLT

Points on the AVLT represent correctly recalled words. The PE-unadjusted model indicated nonsignificant improved performance over time. In contrast, the PE-adjusted model indicated significantly worsening performance over time (−.30 points, −.61 points). The

Table 1
Descriptive Statistics

(a) Sample demographics at baseline

Group	*n*	Baseline age	Baseline ANART*	Education
CU	809	73.70 (6.79)	10.16 (7.94)	16.38 (2.64)
MCI	381	73.07 (7.34)	13.42 (9.64)	15.98 (2.83)

(b) Mean and standard deviation of neuropsychological test scores at each visit

		M (*SD*) at baseline					
	n	LM	AVLT	BNT	CF	TA	TB
CU	809	10.79 (4.21)	7.19 (3.79)	27.89 (2.36)	20.07 (5.25)	34.22 (10.84)	85.76 (38.74)
MCI	381	5.18 (3.61)	2.02 (2.66)	25.50 (4.05)	15.97 (4.81)	45.79 (21.81)	130.64 (70.01)
		M (*SD*) at 12-month follow-up					
	n	LM	AVLT	BNT	CF	TA	TB
CU	735	12.06 (4.54)	7.37 (4.07)	28.39 (1.97)	20.28 (5.17)	33.35 (10.40)	84.70 (41.84)
MCI	381	4.42 (3.88)	1.47 (2.64)	25.67 (4.80)	15.38 (5.51)	46.94 (23.42)	136.56 (77.10)
		M (*SD*) at 24-month follow-up					
	n	LM	AVLT	BNT	CF	TA	TB
CU	691	12.97 (4.18)	7.72 (4.13)	28.44 (2.11)	20.53 (5.31)	32.14 (10.87)	82.77 (40.35)
MCI	241	4.75 (3.95)	1.36 (2.29)	25.95 (4.43)	15.32 (4.93)	44.42 (24.12)	119.78 (64.30)

Note. Presents the average (standard deviation) age, education, and scores on cognitive testing for the subsample of participants who were cognitively unimpaired (CU) at baseline and those diagnosed with mild cognitive impairment (MCI) at baseline. Significant differences ($p < .05$) between subsamples (MCI vs. CU) are designated with a "*." The American National Adult Reading Test (ANART) was given at baseline only, and the provided scores are the total number of errors on the test. The listed ANART errors correspond to an estimated verbal IQ of 118 (*SD* = 9.16) and 117 (*SD* = 9.71) for the CU and MCI samples, respectively (Grober et al., 1991). The remaining cognitive tests were completed at baseline, a 12-month follow-up, and a 24-month follow-up. Means (standard deviations) are presented for each cognitive measure. The number of participants within each group at each visit is presented in the leftmost column. LM = Logical Memory; AVLT = Rey Auditory Verbal Learning Task; BNT = Boston Naming Test; CF = Category Fluency; TA = Trails A; TB = Trails B.

PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 24-month follow-up ([−.73, 1.4] vs. [−.90, −3.2]).

BNT

Points on the BNT represent pictures correctly named. The PE-unadjusted model indicated nonsignificant improvement in scores over time. The PE-adjusted model showed nonsignificant improvement at the 12-month follow-up but significantly worse scores at the 24-month follow-up (−.56 points).

Category Fluency

Points on category fluency represent words generated. The PE-unadjusted model indicated nonsignificant improvement in scores

Table 2
Estimates for Generalized Estimating Equation Models Within the Subsample Diagnosed as Cognitively Unimpaired at Baseline

	12-month follow-up			24-month follow-up		
Measure	Unadjusted age effect	Practice effect	Adjusted age effect	Unadjusted age effect	Practice effect	Adjusted age effect
LM	+1.74* **[1.4, 2.1]** $p < .001$	+2.28* [1.5, 3.1] $p < .001$	−.54* **[−.84, −.24]** $p < .001$	+1.13* [**.83, 1.4**] $p < .001$	+3.18* [2.2,4.2] $p < .001$	−2.0* **[−2.3, −1.7]** $p < .001$
AVLT	+.74 [−.11, 1.6] $p = .096$	+.98* [.38, 1.6] $p = .001$	−.30* [−.57, −.02] $p = .033$	+.32 **[−.73, 1.4]** $p = .181$	+.93* [.02,1.8] $p = .045$	−.61* **[−.90, −3.2]** $p < .001$
BNT	+.44 [−.12, 1.0] $p = .258$	+.46* [.09, .84] $p = .014$	+.02 [−.13, .16] $p = .818$	+.08 [−.46, .63] $p = .086$	+.65* [.14, 1.2] $p = .012$	−.56* [−.73, −.39] $p < .001$
CF	+.72 [−.44, 1.9] $p = .200$	+.96* [.07, 1.9] $p = .035$	−.19 [−.54, .16] $p = .279$	+.40* [.05, .75] $p = .025$	+1.05 [−.25, 2.4] $p = .113$	−.60* [−.97, −.23] $p = .002$
Trails A	+2.72 **[−.11, 5.5]** $p = .076$	+3.81* [1.8, 5.8] $p < .001$	−.99* **[−1.8, −.22]** $p = .012$	+1.29* [**.32, 2.3**] $p = .011$	+3.40* [.74,6.1] $p = .012$	−2.23* **[−3.1, −1.4]** $p < .001$
Trails B	+11.6 [−1.8, 25.0] $p = .698$	+12.94* [5.5, 20.4] $p < .001$	−1.42 [−4.4, 1.6] $p = .353$	+6.87 **[−.07, 13.8]** $p = .056$	+9.59* [−.15,19.3] $p = .054$	−4.0* **[−7.1, −.82]** $p = .014$

Note. Presents unstandardized regression coefficients [confidence intervals] for age effects and practice effects (PEs) in 12 generalized estimating equation (GEE) models. A positive number indicates an improvement in scores as compared to baseline; Trials A and Trails B have been reverse-scored for ease of interpretation. The "unadjusted age effect" columns present results from the PE-unadjusted GEE models and demonstrate how scores at the 12-month and 24-month follow-up visits differ from baseline. The remaining columns present results from the PE-adjusted GEE models. The "practice effect" column provides the PE estimate, and the "adjusted age effect" presents the change over time in scores after correcting for PEs. Estimates significant at $p < .05$ have indicated with an "*." Bold values indicate nonoverlapping CI between the associated unadjusted age effect and the adjusted age effect. LM = Logical Memory; AVLT = Rey Auditory Verbal Learning Task; BNT = Boston Naming Task; CF = Category Fluency; CI = confidence interval.

Table 3
Practice Effect and Time Estimates for Generalized Estimating Equations Within a Subsample Diagnosed With Mild Cognitive Impairment at Baseline

	12-month follow-up			24-month follow-up		
Measure	Unadjusted age effect	Practice effect	Adjusted age effect	Unadjusted age effect	Practice effect	Adjusted age effect
LM	+.79 [−1.9, 3.5] p = .140	+1.50* [.84, 2.2] p < .001	−.72 [−1.1, −.35] p < .001	+2.4 **[−.11, 4.9]** p = .062	+3.50* [1.9, 5.1] p < .001	−1.0* **[−1.5, −.52]** p < .001
AVLT	+.07 **[−.21, .35]** p = .224	+.62* [.15, 1.1] p = .010	−.52 **[−.80, −.24]** p < .001	−.09 **[−.44, .26]** p = .252	+.80 [−.16, 1.8] p = .103	−.89 **[−1.3, −.53]** p < .001
BNT	+1.3 [−.30, 2.8] p = .176	+1.01 [.26, 1.8] p = .009	+.24 [−.14, .62] p = .213	−.58 [−2.5, 1.3] p = .522	−.92 [−2.7, .86] p = .309	+.36 [−.09, .81] p = .119
CF	+.99* [−9.6, 2.9] p = .049	+1.50 [.62, 2.4] p = .001	−.54* [−1.0, −.03] p = .037	−2.0 [−4.4, .49] p = .165	−.80 [−3.0, 1.4] p = .470	−1.01* [−1.6, −.38] p = .002
Trails A	+5.96 [−1.6, 13.5] p = .644	+6.68* [2.2, 11.1] p = .003	−.57 [−2.8, 1.7] p = .622	−.92 [−4.0, 2.2] p = .116	−.90 [−15.6, 13.8] p = .905	.00 [−2.9, 2.9] p = .998
Trails B	+26.30 [−5.2, 57.8] p = .301	+35.34* [21.7, 49.0] p < .001	−3.24 [−10.0, 3.6] p = .350	+30.01 [−1.2, 61.2] p = .103	+30.66* [2.2, 59.1] p = .035	+1.24 [−6.9, 9.3] p = .764

Note. Presents unstandardized regression coefficients [confidence intervals] for age effects and practice effects (PEs) in 12 generalized estimating equation (GEE) models. A positive number indicates an improvement in scores as compared to baseline; Trials A and Trails B have been reverse-scored for ease of interpretation. The "unadjusted age effect" columns present results from the PE-unadjusted GEE models and demonstrate how scores at the 12-month and 24-month follow-up visits differ from baseline. The remaining columns present results from the PE-adjusted GEE models. The "practice effect" column provides the PE estimate, and the "adjusted age effect" presents the change over time in scores after correcting for PEs. Estimates significant at p < .05 have indicated with an "*." Bold values indicate nonoverlapping CI between the associated unadjusted age effect and the adjusted age effect. LM = Logical Memory; AVLT = Rey Auditory Verbal Learning Task; BNT = Boston Naming Task; CF = Category Fluency; CI = confidence interval.

at the 12-month follow-up and significant improvement in scores at the 24-month follow-up (+.40 points). In the PE-adjusted model, there was nonsignificant worsening of scores at the 12-month follow-up and significant worsening of scores at the 24-month visit (−.60 points).

Trails A

Trails A scores indicate seconds to complete the task. Data have been reverse-scored so that higher values indicate better performance. The PE-unadjusted model indicated nonsignificant improvement in scores at the 12-month follow-up and significant improvement in scores at the 24-month follow-up (+1.29 s). In contrast, when adjusting for PEs, there was nonsignificant worsening of scores at the 12-month visit and significant worsening of scores at the 24-month follow-up (−2.23 s). The PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 12-month follow-up ([−.11, 5.5] vs. [−1.8, −.22]) and the 24-month follow-up ([.32, 2.3] vs. [−3.1, −1.4]).

Trails B

Trails B scores indicate seconds to complete the task. Data have been reverse-scored so that higher values indicate better performance. The PE-unadjusted model indicated nonsignificant improvement in scores at the 12-month follow-up and at the 24-month follow-up. The PE-adjusted model indicated nonsignificant worsening in scores at the 12-month follow-up and a significant worsening of scores at the 24-month follow-up (−4.0 s). The PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 12-month follow-up ([−.07, 13.8] vs. [−7.1, −.82]).

Cognitive Trajectories With and Without PE Adjustment in the MCI Group

Within the MCI group, there were notable differences in change-over-time estimates between the PE-adjusted model and the PE-unadjusted models. Figure 3 displays the PE-adjusted and PE-unadjusted trajectories for all cognitive measure among these participants.

Logical Memory

In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up and at the 24-month follow-up. The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up and a significant worsening of scores at the 24-month follow-up (−1.0 points). The PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 24-month follow-up ([−.11, 4.9] vs. [−1.5, −.52]).

AVLT

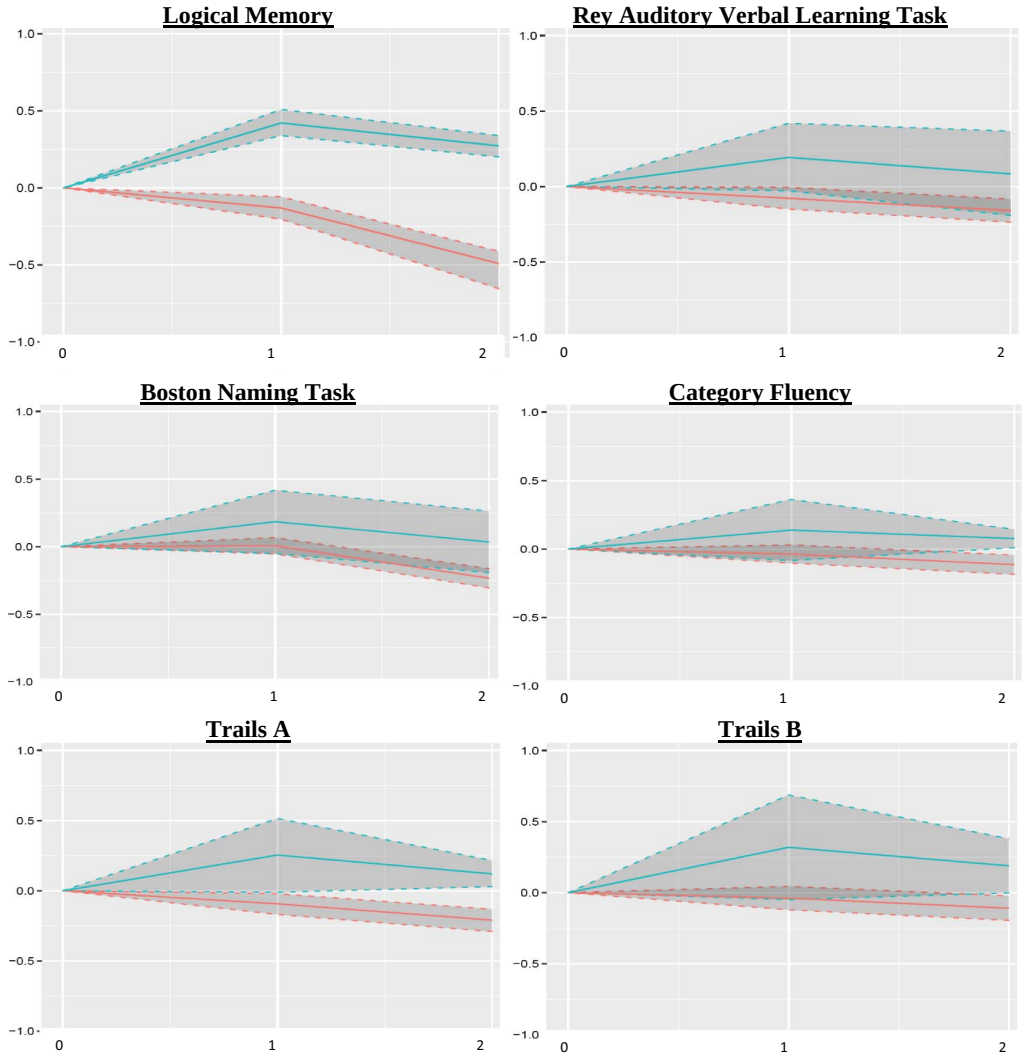
In the PE-unadjusted model, there was nonsignificant change in scores at the 12-month follow-up and at the 24-month follow-up. The PE-adjusted model showed a nonsignificant worsening of scores at the 12-month follow-up and a significant worsening of scores at the 24-month follow-up (−.89 points). The PE-unadjusted and the PE-adjusted age effects had nonoverlapping confidence intervals at the 12-month follow-up ([−.21, .35] vs. [−.80, −.24]) and the 24-month follow-up ([−.44, .26] vs. [−.13, −.53]).

BNT

In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up and nonsignificant

Figure 2

Expected Cognitive Scores Among Participants Who Were Cognitively Unimpaired at Baseline



Note. The *y*-axis of each graph presents standardized scores for all six cognitive measures. The *x*-axis indicates the baseline (0), 12-month follow-up (1), and 24-month follow-up (2) visits. The blue line provides estimates from the practice effect-unadjusted generalized estimating equation model; the red line presents estimates from the practice effect-adjusted model. Dashed lines represent 95% confidence intervals for the practice effect-unadjusted models (blue line) and the practice effect-adjusted models (red lines). Lines with negative slopes indicate that participants' scores are worsening as they age. All participants in these models were diagnosed as cognitively unimpaired at baseline. Trails A and Trials B were reverse-scored to ease interpretation. Raw scores are not presented as differences in the scales of the cognitive tests made visualization unclear. See the online article for the color version of this figure.

worsening of scores at the 24-month follow-up. The PE-adjusted model showed nonsignificant improvement in scores at the 12-month follow-up and at the 24-month follow-up.

Category Fluency

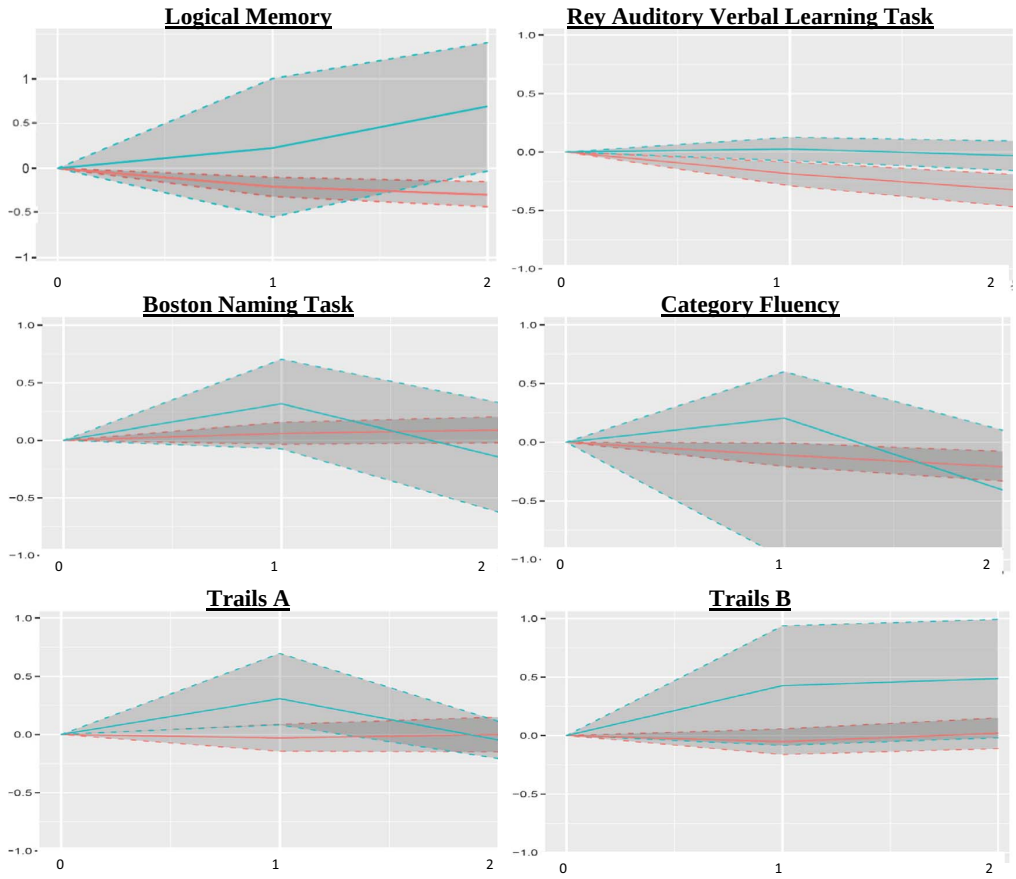
In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up and nonsignificant worsening of scores at the 24-month follow-up. The PE-adjusted model showed significant worsening of scores at the 12-month follow-up (−.54 points) and at the 24-month follow-up (−1.01 points).

Trails A

In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up and nonsignificant worsening of scores at the 24-month follow-up. The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up and a near-zero change at the 24-month follow-up.

Figure 3

Expected Cognitive Scores Among Participants Diagnosed With Mild Cognitive Impairment at Baseline



Note. The *y*-axis of each graph presents standardized scores for all six cognitive measures. The *x*-axis indicates the baseline (0), 12-month follow-up (1), and 24-month follow-up (2). The blue line provides estimates from the practice effect-unadjusted generalized estimating equation model; the red line presents estimates from the practice effect-adjusted model. Dashed lines represent 95% confidence intervals for the practice effect-unadjusted models (blue line) and the practice effect-adjusted models (red lines). Lines with negative slopes indicate that participants' scores are worsening as they age. All participants in these models were diagnosed with mild cognitive impairment at baseline. Trails A and Trials B were reverse-scored to ease interpretation. Raw scores are not presented as differences in the scales of the cognitive tests made visualization unclear. See the online article for the color version of this figure.

Trails B

Within the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up and at the 24-month follow-up. The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up and a nonsignificant improvement in scores at the 24-month follow-up.

Discussion

Using the replacement participants method of PE adjustment, in combination with GEE, we found significant PEs across two follow-up visits at 12-month intervals in baseline CU and MCI groups. The magnitudes of PEs at the 12-month follow-up within the group that was CU at baseline were not consistently larger than those within the group that was MCI at baseline. This is somewhat inconsistent with the PE literature, which typically finds that long-term PEs are larger in those who are CU at baseline (Galvin et al., 2005; Goldberg et al., 2015; Schrijnemaekers et al., 2006). However, in the present study, the PEs within the MCI group had notably larger confidence intervals than those in the CU group. The difference may reflect greater variability in the MCI group or it could be due to the smaller MCI group size (809 vs. 381). Additionally, some of the MCI participants were at floor on memory measures as well as other tasks. It is possible that floor effects impacted these results and were responsible for the high heterogeneity and nonsignificance of the PE estimates among the MCI participants.

At the 24-month follow-up visit the CU group had significant PEs on five of the six measures while the MCI group had only two significant PEs. However, as the magnitude of these PEs were similar between the MCI and the CU groups, it is possible that the smaller sample size of the MCI group reduced the power to detect significant results. With a greater number of MCI participants it is likely that more PEs would remain significant. Of note, the PEs on the Logical Memory measure were significant for both groups at

both timepoints. Performance below impairment cutoffs on Logical Memory is one of the primary criteria for ADNI's MCI diagnosis (Petersen et al., 2010). These results suggest that both the CU and MCI participants are experiencing important PEs that are likely impacting incidence rates of MCI and stability of MCI diagnoses within ADNI (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022; Thomas et al., 2019). Because accounting for PEs with this method lowers scores on all tests, it affects all MCI diagnoses that use cutoff scores (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022), whether impairment on one (Petersen et al., 2010) or more than one measure (Jak et al., 2009) is required. Of note, we did not correct for multiple comparisons within this study and reported all effects, indicating those that were significant with an α of .05. In total, 17 of the 24 PEs were statistically significant at this α level. Prior work has demonstrated that even small PEs can have meaningful downstream effects, specifically on diagnosis of MCI, reversion to CU, and biomarker–cognition concordance (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022).

A goal of this project was to investigate how adjusting for PEs impacts measurement of cognitive trajectories. We completed analyses twice, modeling change over time at 12-month intervals with and without considering PEs. When PEs were not included in the models, scores tended to increase or stay the same over time. If this trend accurately reflected cognition, it would mean that these older adults were improving their cognitive ability over 2 years. While possible, this interpretation is highly unlikely given that the norm for adults in this age range is for cognitive decline over time (Salthouse, 2010, 2019). Moreover, the parent study, ADNI, recruited participants that were similar to those in AD clinical drug trials and who have a high risk for neurodegeneration (Petersen et al., 2010). In contrast, when including PEs in the models, there was worsening performance across visits that in many cases was significantly different from the models that did not adjust for PEs. In some instances, the age effect was nonsignificant in the PE-adjusted model. However, the confidence intervals were nonoverlapping between PE-unadjusted and PE-adjusted age effect estimates. This suggests that there was a significance difference in change over time when PEs were included in the model, even if the PE-adjusted age-effect estimate was nonsignificant. Taken together, we interpret these results to indicate that adjusting for PEs led to results that more accurately represent the expected cognitive change.

The confidence intervals in Figures 2 and 3 further demonstrate how PE adjustment impacts the precision of longitudinal analyses. The confidence intervals of the unadjusted estimates are notably wider than those for the PE-adjusted estimates. For example, the confidence intervals for the category fluency estimates within the MCI baseline sample (Figure 3) all suggest that the time trends are nonsignificantly different from zero. However, the PE-unadjusted confidence intervals appear more than twice as large as those in the PE-adjusted estimates. This pattern is similar in many of the other comparisons. We interpret this pattern as secondary support for our hypothesis that PE-adjusted scores lead to more accurate estimates of time trends. From a broader perspective, this means that failing to address PEs leads to inaccurate interpretation of cognitive performance. These results are consistent with prior research claiming that PE adjustment can be viewed as a data correction tool that improves the accuracy of cognitive data and diagnoses (e.g., earlier diagnosis, more stable diagnoses; Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). However, it is important to note that, in practice, other methods of accounting for PEs do not alter diagnosis; they simply describe PEs or cognitive trajectories. In contrast, with the participant-replacement method, cognitive scores are adjusted downward, which in turn, leads to earlier diagnosis of disorders such as MCI because more individuals drop below the impairment cutoff. Clinically, the importance of detecting a diagnosis that involves impaired functioning as early as possible is of obvious importance. In research, as diagnostic groups and cognitive scores are basic components of outcome measures, these small PE adjustments can have significant downstream effects on everything from the utility of biomarkers to study duration (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022).

PEs have been shown to weaken over time in some studies, leading many to conclude that consideration of PEs is less important at subsequent follow-up visits (Goldberg et al., 2015; Heilbronner et al., 2010; Machulda et al., 2017; Stricker et al., 2020; Vivot et al., 2016). However, research supporting this claim focuses on methods that almost always define PEs as increases in scores, which means they are difficult, if not impossible, to observe in the presence of overall (e.g., age-related) decline (Calamia et al., 2012; Elman et al., 2018; Goldberg et al., 2015; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022). Salthouse (2019) noted that failure to consider how PEs and age-related decline interact over time is a significant barrier to cognitive research and one of the principal contributors to the discrepancy between cross-sectional and longitudinal research findings (Salthouse, 2019). He recommended a quasi-longitudinal approach that incorporates a cross-sectional group of adults who are similar to returning participants, which is very similar to the replacement method of PE adjustment (Salthouse, 2019). Using a participant replacement method, we showed that PEs do not uniformly decrease across visits. In fact, the magnitude of two of the PEs that remained significant across visits increased over time within the CU group, and one increased within the MCI group. Additionally, most of the PEs within both groups remained fairly stable across time, although many of the PEs in the MCI group were nonsignificant at the 24-month follow-up. These results indicate that PEs may have a different trajectory than what is generally considered, and they are clearly not uniform across different cognitive tests.

We believe that many PE definitions—which almost always restrict PEs to an observed increase in test scores—are less accurate in groups where pathology- or age-related decline is expected. For example, consistent with the widely held view of cognitive PEs, a recent study of older adults with multiple follow-ups over a more than 4-year period found that PEs largely leveled out after the first follow-up assessment (Stricker et al., 2020). As is also common (Holm et al., 2022), they defined PEs as increases in scores and they

controlled for age in their models. They concluded that PEs were worth considering at all visits but that the most important was the 12-month follow-up. We generally agree with the conclusions of this extremely well-done study (Stricker et al., 2020). However, as with many studies, we think that this approach likely underestimates PEs, which in turn may alter the interpretation of cognitive change (Salthouse, 2019). Defining PEs solely on the basis of improved scores limits their detection when performance declines, as is expected in aging and neurodegenerative diseases. PE models that include age as a covariate do not account for how aging may impact PEs. We have shown that when co-occurring with age-related decline, PEs may exist even when the actual scores decrease. The advantage of a participant replacement method is that accounting for age-related decline is built into the calculation of PEs, which makes it possible to show PEs even when scores worsen over time. It is also worth noting that this feature is not restricted to samples of very old adults. For example, Elman et al. demonstrated the same phenomenon in a 6-year follow-up of adults who were in their mid-50s at baseline.

In a clinical setting, neuropsychologists do not have the potential benefit of replacement participants as may be the case in research. Although this method is not applicable within clinics, our results may improve how clinicians consider longitudinal change. We found that PEs do not uniformly decline after the first visit and posit that it is important to consider PEs even if scores do not increase at follow-up. We recommend that clinicians be wary of underestimating change if their at-risk patients appear to minimally decline at follow-up, particularly if they or their informant report subjective changes. Our work also suggests the potential value of developing PE norms at common clinical follow-up intervals (e.g., 6 months, 1 year).

As the field of clinical neuropsychology evaluates the current state of normative data (Byrd & Rivera-Mindt, 2022), and while there is a strong push for repeated assessments, particularly via computerized testing (Jutten et al., 2022; Stricker et al., 2020), the results of the present study strongly suggest that the field would benefit from normative data for PEs over multiple retest visits. In particular, it may be beneficial to develop predicted PEs for specific tests in a diverse range of individuals retested at clinically relevant intervals across multiple retest visits (e.g., every 6 or 12 months).

Generalizability of Cognitive Practice Effects: Yes or No?

Generalizability is an important issue regarding research results and restrictions to generalizability are usually noted as limitations. We specifically did not include a detailed discussion of PE generalizability in the limitations section (see below) because we believe this concern reflects a key misunderstanding of the replacement-participants method and of the nature of cognitive PEs. At a very broad level, our results are generalizable; PEs are found with the participant replacement method across different retest intervals and cognitive measures in very different samples. Significant PEs were observed in a Swedish population-based sample with baseline ages of 24–80 and a 5-year follow-up; a sample of community-dwelling U.S. men ages 51–60 with a 6-year follow-up; and the clinic-based sample of adults in the present study at average age of 74 at baseline with 1-year follow-ups (Elman et al., 2018; Rönnlund et al., 2005; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022). Similarly, the idea that accounting for PEs with this method results in earlier diagnosis of MCI is generalizable; this result has been demonstrated in the community-based Vietnam Era Twin Study of Aging (VETSA) sample and the older, clinic-based ADNI sample (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022).

However, the precise magnitude of PEs (e.g., exact percent change in MCI status) is not generalizable from one sample to another. PEs will almost always vary based on age or other demographic characteristics, the specific cognitive measures, and the different retest intervals across studies (Glisky et al., 2022; Kremen et al., 2022). Given this heterogeneity in PE magnitudes, we discourage the generalization of PEs based on research studies. This viewpoint is implicitly echoed in the results of studies using other approaches to long-term PEs, including RCIs and standardized regression-based approaches (Chelune et al., 1993; Duff, 2012; Duff & Hammers, 2022). For example, use of a meta-analysis to provide specific magnitudes of PEs generalized poorly to a separate sample (Duff & Hammers, 2022).

The creation of truly generalizable PEs may be possible but doing so would be a major undertaking on par with the creation of a normative data set based on demographically inclusive samples with a wide range of ages, retest intervals, and education. Generalizability of these PE norms would then still be limited to only the age groups, specific tests, and the particular retest intervals that were included in the normative data samples. If any of these three sets of features differ, there should be no expectation that the magnitude of PEs should be generalizable.

A key feature of the replacement-participants method is that it does not require generalizability for the results to be meaningful and valid. By design, the replacement participants are always precisely matched with returnees on demographic characteristics, cognitive measures, and duration of retest intervals specific to that study. Thus, the magnitude of PEs may vary across studies, but they are valid within any individual study and can impact downstream analyses. For example, within the present study, PE-adjustment led to a more accurate characterization of cognitive aging trajectories, as compared to unadjusted scores. Similarly, as compared to PE-unadjusted scores, other studies using the method have demonstrated that PE adjustment leads to earlier detection of progression from CU to MCI status and better concordance between cognitive status and biomarker profiles (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022).

Clinical Versus Statistical Significance

Results from the present study and prior studies using the replacement-participants method show that the method results in clinically meaningful differences. As shown elsewhere, small—even nonsignificant—raw score differences can be enough to move people across the threshold for diagnosis of MCI, resulting in a higher number of people diagnosed with MCI (Elman et al., 2018; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022). For example, within one study of PEs in ADNI, there was a 19% increases in MCI diagnoses after adjusting for PEs. In a separate study in the VETSA, PE adjustment led to a 100% increase (doubling) in MCI diagnoses. Those

increases translated to 20 additional ADNI and 45 additional VETSA individuals being diagnosed with MCI one visit earlier than they would have using PE-unadjusted scores. As neuropsychological assessments typically have at least a 12-month retest interval, diagnosis of MCI one visit earlier can have very meaningful impact for those individuals. Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al. (2022) also established the validity of these diagnoses based on stronger concordance with AD biomarkers for diagnoses based on PE-adjusted compared with PE-unadjusted scores. The earlier detection of MCI by accounting for PEs with this method would also shorten study duration and increase statistical power in a clinical trial. Using the A4 trial, a major Alzheimer's clinical trial, one study calculated that these changes could result in a savings of study costs of well over $5 million (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022). These results demonstrate that the method does increase precision enough to make a clinically meaningful, and not just a statistically significant, difference.

Strengths and Limitations

Participants for this project were drawn from ADNI, which consisted primarily of highly educated, healthy, white individuals who typically present to memory clinics (Petersen et al., 2010). This sample is not representative of the U.S. population. As noted above, we explicitly recommend against generalizing PEs for use in other samples.

The retention rates differed between the CU and the MCI groups (85% vs. 63%). Attrition has been shown to impact PE estimates (Rönnlund et al., 2005; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022) and our GEE models did not include attrition effects. As the MCI group had a higher attrition rate, it is possible that the impact of the attrition rate on the PEs was larger in the MCI group than in the CU group. Future studies should include attrition rates in their models.

The replacement participant method generates a group mean effect for each cognitive measure. This can be viewed as a limitation by some as many methods focus on individual PEs with the goal of predicting who will progress to dementia. However, those studies also are unable to alter how and when a diagnosis is made, a feature of the replacement method's group level PE. As noted in an overview of selected PE studies of aging, there is no one-size-fits-all solution and there is no single best PE method; it is simply that different methods address different questions (Kremen et al., 2022). This represents a limitation in that the method has not been shown to be readily applicable to smaller studies that have, say, 30, 50, or 100 participants in total. However, to our knowledge, no study has provided a comprehensive investigation of the number of necessary replacement participants, and it may be possible to apply this method to smaller samples. In a prior study, there were significant PEs and a significant increase in MCI cases with about 170 replacement subjects, leading the authors to estimate that 150–200 matched replacements would probably be adequate (Elman et al., 2018). Additionally, we point to cognitive aging and Alzheimer's-related research as one example for which there are dozens of studies that are easily large enough to incorporate the replacement method. It is the large studies that tend to be the most important and the ones with the highest impact. The rapid growth of "big science" and "big data" has also demonstrated that large samples can detect things that small studies cannot, not to mention the fact that large samples are more likely to generate reproducible results. Moreover, as noted above, prior results show that despite the added cost of replacements, replacements can still be cost effective (Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022).

Summary

In sum, using the replacement-participants method we found that PEs on several measures did not level out after the first follow-up visit. This finding is in contrast to the predominant view of PEs in neuropsychology. Indeed, PEs for some—particularly episodic memory measures—actually increased at the second (24-month) follow-up. It is possible that PEs will stabilize and decline at subsequent visits but that is unknown at this time. Future studies using this method should investigate measures across a longer time frame. Additionally, the ADNI MCI sample consists primarily of individuals with memory concerns and/or who have been diagnosed with amnestic MCI (Eppig et al., 2017; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eglit, et al., 2022; Sanderson-Cimino, Elman, Tu, Gross, Panizzon, Gustavson, Bondi, Edmonds, Eppig, et al., 2022; Thomas et al., 2019). Future studies with larger sample sizes may benefit from conducting subanalyses delineated by MCI subtype. However, making maximal use of PEs is probably most relevant and most important in individuals who are CU in order to foster detection of progression to MCI at the earliest possible time.

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). 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. *Alzheimer's & Dementia*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Byrd, D. A., & Rivera-Mindt, M. G. (2022). Neuropsychology's race problem does not begin or end with demographically adjusted norms. *Nature Reviews. Neurology*, *18*(3), 125–126. https://doi.org/10.1038/s41582-021-00607-4

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Chelune, G. J., Bornstein, R. A., & Prifitera, A. (1990). The Wechsler memory scale—Revised. In P. McReynolds (Ed.), *Advances in psychological assessment* (pp. 65–99). Springer. https://doi.org/10.1007/978-1-4613-0555-2_3

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*(1), 41–52. https://doi.org/10.1037/0894-4105.7.1.41

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical Neuropsychologist*, *28*(5), 714–725. https://doi.org/10.1080/13854046.2014.920923

Duff, K., Foster, N. L., & Hoffman, J. M. (2014). Practice effects and amyloid deposition: Preliminary data on a method for enriching samples in clinical trials. *Alzheimer Disease and Associated Disorders*, *28*(3), 247–252. https://doi.org/10.1097/WAD.0000000000000021

Duff, K., & Hammers, D. B. (2022). Practice effects in mild cognitive impairment: A validation of Calamia et al. (2012). *The Clinical Neuropsychologist*, *36*(3), 571–583. https://doi.org/10.1080/13854046.2020.1781933

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., Schultz, S. K., Paulsen, J. S., Petersen, R. C., & McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry*, *19*(11), 932–939. https://doi.org/10.1097/JGP.0b013e318209dd3a

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., Gustavson, D. E., Franz, C. E., Hatton, S. N., Jacobson, K. C., Toomey, R., McKenzie, R., Xian, H., Lyons, M. J., & Kremen, W. S. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*(1), 372–381. https://doi.org/10.1016/j.dadm.2018.04.003

Eppig, J. S., Edmonds, E. C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M. W., & the Alzheimer's Disease Neuroimaging Initiative. (2017). Statistically derived subtypes and associations with cerebrospinal fluid and genetic biomarkers in mild cognitive impairment: A latent profile analysis. *Journal of the International Neuropsychological Society*, *23*(7), 564–576. https://doi.org/10.1017/S135561771700039X

Finkel, D., Reynolds, C. A., McArdle, J. J., Gatz, M., & Pedersen, N. L. (2003). Latent growth curve analyses of accelerating decline in cognitive abilities in late adulthood. *Developmental Psychology*, *39*(3), 535–550. https://doi.org/10.1037/0012-1649.39.3.535

Galvin, J. E., Powlishta, K. K., Wilkins, K., McKeel, D. W., Jr., Xiong, C., Grant, E., Storandt, M., & Morris, J. C. (2005). Predictors of preclinical Alzheimer disease and dementia: A clinicopathologic study. *Archives of Neurology*, *62*(5), 758–765. https://doi.org/10.1001/archneur.62.5.758

Glisky, E. L., Woolverton, C. B., McVeigh, K. S., & Grilli, M. D. (2022). Episodic memory and executive function are differentially affected by retests but similarly affected by age in a longitudinal study of normally-aging older adults. *Frontiers in Aging Neuroscience*, *14*, Article 863942. https://doi.org/10.3389/fnagi.2022.863942

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 103–111. https://doi.org/10.1016/j.dadm.2014.11.003

Grober, E., Sliwinski, M., & Korey, S. R. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, *13*(6), 933–949. https://doi.org/10.1080/01688639108405109

Gross, A. L., Anderson, L., & Chu, N. (2017). Do people with Alzheimer's disease improve with repeated testing? Unpacking the role of content and context in retest effects. *Alzheimer's & Dementia*, *13*(7), 473–P474. https://doi.org/10.1016/j.jalz.2017.06.497

Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., Sisco, S., Skinner, J., Schneider, B. C., & Manly, J. J. (2015). Predictors of retest effects in a longitudinal study of cognitive aging in a diverse community-based sample. *Journal of the International Neuropsychological Society*, *21*(7), 506–518. https://doi.org/10.1017/S1355617715000508

Gross, A. L., Chu, N., Anderson, L., Glymour, M. M., Jones, R. N., Diseases, C. A. M., & the Coalition Against Major Diseases. (2018). Do people with Alzheimer's disease improve with repeated testing? Unpacking the role of content and context in retest effects. *Age and Ageing*, *47*(6), 866–871. https://doi.org/10.1093/ageing/afy136

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). 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. *The Clinical Neuropsychologist*, *24*(8), 1267–1278. https://doi.org/10.1080/13854046.2010.526785

Holm, S. P., Wolfer, A. M., Pointeau, G. H. S., Lipsmeier, F., & Lindemann, M. (2022). Practice effects in performance outcome measures in patients living with neurologic disorders—A systematic review. *Heliyon*, *8*(8), Article e10259. https://doi.org/10.1016/j.heliyon.2022.e10259

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., & Delis, D. C. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *The American Journal of Geriatric Psychiatry*, *17*(5), 368–375. https://doi.org/10.1097/JGP.0b013e31819431d5

Jutten, R. J., Thompson, L., Sikkes, S. A., Maruff, P., Molinuevo, J. L., Zetterberg, H., Alber, J., Faust, D., Gauthier, S., Gold, M., Harrison, J., Lee, A. K. W., & Snyder, P. J. (2022). A neuropsychological perspective on defining cognitive impairment in the clinical study of Alzheimer's disease: Towards a more continuous approach. *Journal of Alzheimer's Disease*, *86*(2), 511–524. https://doi.org/10.3233/JAD-215098

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test*. Springer.

Kremen, W. S., Nation, D., & Nyberg, L. (2022). Featured research topic editorial: The importance of cognitive practice effects in aging neuroscience. *Frontiers in Aging Neuroscience*, *14*, 1079021. https://doi.org/10.3389/fnagi.2022.1079021

Lezak, M. D., Howieson, D. B., Loring, D. W., & Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press.

Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., Vemuri, P., Lowe, V. J., Jack, C. R., Jr., & Petersen, R. C. (2017). Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. *The Clinical Neuropsychologist*, *31*(1), 99–117. https://doi.org/10.1080/13854046.2016.1241303

Manly, J. J., Tang, M. X., Schupf, N., Stern, Y., Vonsattel, J. P. G., & Mayeux, R. (2008). Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology*, *63*(4), 494–506. https://doi.org/10.1002/ana.21326

Mathews, M., Abner, E., Kryscio, R., Jicha, G., Cooper, G., Smith, C., Caban-Holt, A., & Schmitt, F. A. (2014). Diagnostic accuracy and practice effects in the National Alzheimer's Coordinating Center Uniform Data Set neuropsychological battery. *Alzheimer's & Dementia*, *10*(6), 675–683. https://doi.org/10.1016/j.jalz.2013.11.007

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jr., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2010). Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Rönnlund, M., Nyberg, L., Bäckman, L., & Nilsson, L.-G. (2005). Stability, growth, and decline in adult life span development of declarative memory: Cross-sectional and longitudinal data from a population-based study. *Psychology and Aging*, *20*(1), 3–18. https://doi.org/10.1037/0882-7974.20.1.3

Saloner, R., Casaletto, K. B., Marx, G., Dutt, S., Vanden Bussche, A. B., You, M., Fox, E., Stiver, J., & Kramer, J. H. (2018). Performance on a 1-week delayed recall task is associated with medial temporal lobe structures in neurologically normal older adults. *The Clinical Neuropsychologist*, *32*(3), 456–467. https://doi.org/10.1080/13854046.2017.1370134

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, *16*(5), 754–760. https://doi.org/10.1017/S1355617710000706

Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and Aging*, *34*(1), 17–24. https://doi.org/10.1037/pag0000288

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., Bondi, M. W., Edmonds, E. C., Eglit, G. M. L., Eppig, J. S., Franz, C. E., Jak, A. J., Lyons, M. J., Thomas, K. R., Williams, M. E., Kremen, W. S., & Alzheimer's Disease Neuroimaging Initiative. (2022). Cognitive practice effects delay diagnosis of MCI; Implications for clinical trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*. *8*(1), Article e12228. https://doi.org/10.1002/trc2.12228

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., Bondi, M. W., Edmonds, E. C., Eppig, J. S., Franz, C. E., Jak, A. J., Lyons, M. J., Thomas, K. R., Williams, M. E., & Kremen, W. S. (2022). Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments. *Frontiers in Aging Neuroscience*, *14*, Article 847315. https://doi.org/10.3389/fnagi.2022.847315

Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook* (Vol. 17). Western Psychological Services Los Angeles.

Schrijnemaekers, A. M., de Jager, C. A., Hogervorst, E., & Budge, M. M. (2006). Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. *Journal of Clinical and Experimental Neuropsychology*, *28*(3), 438–455. https://doi.org/10.1080/13803390590935462

Stricker, N. H., Lundt, E. S., Alden, E. C., Albertson, S. M., Machulda, M. M., Kremers, W. K., Knopman, D. S., Petersen, R. C., & Mielke, M. M. (2020). Longitudinal comparison of in clinic and at home administration of the cogstate brief battery and demonstrated practice effects in the Mayo Clinic Study of Aging. *The Journal of Prevention of Alzheimer's Disease*, *7*(1), 21–28. https://doi.org/10.14283/jpad.2019.35

Tang, W., He, H., & Tu, X. M. (2012). *Applied categorical and count data analysis*. CRC Press. https://doi.org/10.1201/b12123

Taylor, K. I., Salmon, D. P., Rice, V. A., Bondi, M. W., Hill, L. R., Ernesto, C. R., & Butters, N. (1996). Longitudinal examination of American National Adult Reading Test (AMNART) performance in dementia of the Alzheimer type (DAT): Validation and correction based on degree of cognitive decline. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 883–891. https://doi.org/10.1080/01688639608408309

Team, R. C. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Thomas, K. R., Bangen, K. J., Weigand, A. J., Edmonds, E. C., Wong, C. G., Cooper, S., Delano-Wood, L., Bondi, M. W., & the Alzheimer's Disease Neuroimaging Initiative. (2020). Objective subtle cognitive difficulties predict future amyloid accumulation and neurodegeneration. *Neurology*, *94*(4), e397–e406. https://doi.org/10.1212/WNL.0000000000008838

Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., Jak, A. J., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Edland, S. D., Bondi, M. W., & the Alzheimer's Disease Neuroimaging Initiative. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia*, *15*(10), 1322–1332. https://doi.org/10.1016/j.jalz.2019.06.4948

Thorndike, E. L. (1913). *An introduction to the theory of mental and social measurements*. Teacher's College, Columbia University.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

Vivot, A., Power, M. C., Glymour, M. M., Mayeda, E. R., Benitez, A., Spiro, A., III, Manly, J. J., Proust-Lima, C., Dufouil, C., & Gross, A. L. (2016). Jump, hop, or skip: Modeling practice effects in studies of determinants of cognitive change in older adults. *American Journal of Epidemiology*, *183*(4), 302–314. https://doi.org/10.1093/aje/kwv212

Received July 2, 2022
Revision received January 4, 2023
Accepted February 8, 2023 ■