UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO STATE UNIVERSITY

A Novel Approach to Cognitive Practice Effects within Aging Cohorts: Earlier Detection of Decline and Change in Diagnostic Status

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Clinical Psychology

by

Mark Sanderson-Cimino

Committee in charge

University of California San Diego

Professor William S. Kremen, Chair
Professor Jeremy Elman
Professor Amy Jak
Professor Xin Tu

San Diego State University

Professor Paul Gilbert
Professor Claire Murphy

2022

The Dissertation of Mark Sanderson-Cimino is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego 2022

San Diego State University 2022

Table Of Contents

List of Figures

List of Tables

List of Graphs

Acknowledgments

Data used in the preparation of this proposal were obtained primarily from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

The ADNI (adni.loni.usc.edu) 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 (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. Participants from the ADNI-1, ADNI-GO, and ADNI-2 cohorts were included. Significant references are also made to the Vietnam Era Twin Study of Aging (VETSA).

Chapter 1, in full, is a reprint of the material as it appears in Alzheimer's & Dementia: Translational Research & Clinical Interventions. Mark Sanderson-Cimino, Jeremy A. Elman, Xin M. Tu, Alden L. Gross, Matthew S. Panizzon, Daniel E. Gustavson, Mark W. Bondi, Emily C. Edmonds, Graham M.L. Eglit, Joel S. Eppig, Carol E. Franz, Amy J. Jak, Michael J. Lyons, Kelsey R. Thomas, McKenna E. Williams, William S. Kremen. *Cognitive Practice Effects Delay Diagnosis; Implications for Clinical Trials.* Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Fronters of Aging Neuroscience. Mark Sanderson-Cimino, Jeremy A. Elman, Xin M. Tu, Alden L. Gross, Matthew

S. Panizzon, Daniel E. Gustavson, Mark W. Bondi, Emily C. Edmonds, Joel S. Eppig, Carol E. Franz, Amy J. Jak, Michael J. Lyons, Kelsey R. Thomas, McKenna E. Williams, and William S. Kremen, PhD. *Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments.* Frontiers in Aging Neuroscience, 2022. **14**. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is under peer review. The dissertation author was the primary investigator and author of this paper.

This research was supported by NIA grants R01 AG050595, R01 AG022381, and F31 AG064834.

Vita

2012 Bachelor of Science, Psychology. University of California San Diego

2019 Master's of Science; Clinical Psychology. San Diego State University

2022 Doctor of Philosophy; Clinical Psychology. San Diego State University/University of California San Diego

Publications

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., ... & 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), e12228.

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. 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*.

Field Of Study

Major Field: Clinical Psychology

Studies in Neuropsychology

Abstract of the dissertation

A Novel Approach to Cognitive Practice Effects within Aging Cohorts: Earlier Detection of Decline and Change in Diagnostic Status4

by

Mark Sanderson-Cimino

Doctor of Philosophy in Clinical Psychology

University of California San Diego, 2022

San Diego State University, 2022

Professor William Kremen, Chair

Cognitive practice effects (PEs) are often ignored and likely delay detection of mild cognitive impairment (MCI). The replacement-participants method of PE adjustment subtracts the scores of demographically-matched, test naïve, replacement participants from returnees' follow-up scores. The aims of this project were to determine whether this PE adjustment: 1) results in earlier MCI detection; 2) improves diagnostic stability, accuracy, and validity based on concordance with Alzheimer's disease biomarkers; and 3) reduces costs of clinical trials.

The method was adapted for studies that did not recruit replacement participants by identifying “pseudo-replacements”, a subset of baseline participants that are demographically-matched to returnees at follow-up. Alzheimer’s Disease Neuroimaging Initiative data were extracted. Study 1 (baseline cognitively unimpaired [CU], N=722) and study 2 (baseline MCI, N=329) calculated PEs at 1-year follow-up. Study 3 added data from a 2-year follow-up (N=809). Biomarkers include CSF tau and beta-amyloid. Cost estimates were based on recent grant budgets and effects of PE adjustment on required sample size in a large clinical trial. Primary analyses included McNemar χ 2 tests, logistic regressions, and generalized estimating equations.

Study 1: PE-adjusted follow-up scores led to significantly greater incident MCI as compared to PE-unadjusted scores (124 vs. 104; +19%). Significantly more MCI participants were biomarker positive when PE-adjusted scores were utilized (173 vs. 152; +14%). Cost estimates showed that replacements could save a large Alzheimer’s disease clinical trial over $5,000,000. Study 2: PE adjustment significantly reduced reversion to CU as compared to PE-unadjusted scores (57 vs. 80; -29%). Study 3: Significant PEs were found at both follow-ups for baseline MCI and CU samples. When adjusting for PEs, participants remained stable over time or declined, as expected in this older sample. Several PEs increased at the second follow-up (range: +9% to +133%).

Adjusting for PEs with the replacement method leads to earlier and more accurate MCI diagnoses, reduces reversion rates, and is cost effective for large-scale studies. Unlike other methods, it also unmasks PEs even when scores worsen over time. The results indicate that PEs warrant greater attention and also suggest the potential value of developing PE norms.

1. Introduction

Assessments of cognitive abilities are used to differentiate those who are aging normally (i.e., cognitively unimpaired; CN) from those who are aging abnormally. Those who have slight cognitive deficits in the presence of minimal to no functional impairment may be diagnosed with mild cognitive impairment (MCI), which is often seen as a prodromal stage for several dementias (Albert et al., 2011; Manly et al., 2008). Those who's cognitive decline has notably affected their functionality can be diagnosed with dementia or major neurocognitive disorder (Albert et al., 2011; J. Eppig et al., 2020; Manly et al., 2008; Thomas et al., 2020).

Repeated cognitive assessments are necessary for accurately determining when individuals are declining and when they transition from CU to MCI and to dementia. However, when individuals complete the same test at multiple timepoints their scores are likely affected by practice effects (PEs). PEs are typically defined as an improvement in performance due to retesting, as opposed to true change in cognition (Heilbronner et al., 2010). Put simply, someone taking a test for the second time will typically have a higher score than if they were taking it for the first time. These effects can occur due to memory of specific stimuli (i.e., content PE) or through increased comfort/knowledge with test taking (i.e., context PE) (Gross, Anderson, & Chu, 2017; Heilbronner et al., 2010) PEs are pervasive; they have been found across multiple cognitive domains and test-retest intervals as long as 7 years in older adults, including those with MCI and even mild Alzheimer's disease (AD) (Elman et al., 2018; Goldberg, Harvey, Wesnes, Snyder, & Schneider, 2015; Gross et al., 2017; Gross et al., 2015).

1.1 The Impact of PEs on Diagnosis of Cognitive Deficits

PEs have a wide-ranging impact on any study or field involving cognitive testing because they mask true cognitive decline and compromise diagnostic accuracy, impairing the separation

of cases (i.e., MCI) and controls (i.e., CN) (Calamia, Markon, & Tranel, 2012; Elman et al., 2018). PEs are a barrier to early diagnosis of cognitive impairments because they may obscure decline in two ways. First, PEs may be equal to or larger than age-related decline or early onset symptoms. If so, performance would remain stable or appear to improve, and an individual would seem to be unaffected by age-related changes (Figure A). In this situation, clinicians might not recognize the need to provide early care, and researchers might fail to identify associations between biomarkers and disease progression. The majority of research on PEs considers only this situation (Calamia et al., 2012; Goldberg et al., 2015). Second, even if scores decline, PEs may still be present, masking a much sharper decline (Figure B). Here individuals may appear more cognitively stable than they actually are. This misunderstanding may prevent researchers from accurately mapping cognitive change and disease progression. It also may lead clinicians to conclude that an individual is more capable of independent living than is safe. The second situation is likely in studies of older adults where normative data indicate that cognition is expected to decline over time (Finkel, Reynolds, McArdle, Gatz, & Pedersen, 2003; Salthouse, 2010, 2019).

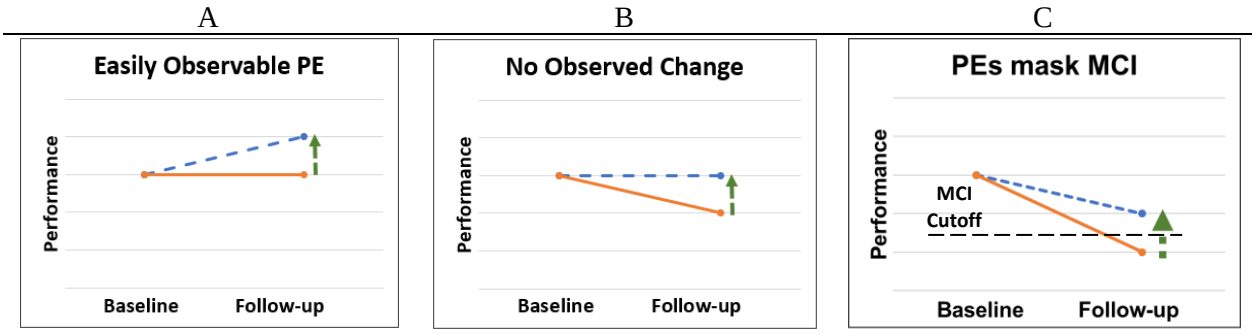


Figure 1: **Theoretical practice effects.** The solid line represents true cognitive ability. The dashed line represents observed performance, which is inflated due to a practice effect (vertical arrow). **1A:** Typically observed practice effect: an individual's observed score increases from baseline to follow-up, demonstrating a typical practice effect. **1B**: Practice effect in the context of cognitive decline. In this scenario, an individual's ability is decreasing overtime. A practice effect still exists but is masked by cognitive decline. As a result, the individual's performance appears to be stable but is actually better than it would have been without previous exposure to the test. **1C**: Practice effects impair detection of MCI. In this situation, an individual has declined below an MCI cutoff. However, practice effects are inflating their score so that they now fall above the MCI cutoff and will be diagnosed as cognitively normal at follow-up.

MCI classification methods, particularly in research, almost always rely on use of cut-off scores to define cognitive impairment (Jak et al., 2009; Winblad et al., 2004). The same cut-off is typically applied at baseline and follow-up visits. If an individual experiences a PE greater than their cognitive decline, then they may be pushed above the threshold for impairment despite having no change or even a slight decline in their actual cognitive ability (Figure C). Among those who are CU at baseline, this means they will be misdiagnosed as CU at follow-up when they should be labeled as MCI. As such, methods that address PEs at follow-up may change MCI classification, increasing diagnostic accuracy. Stated differently, adjusting for PEs may lead to earlier detection of MCI among those who were CU at their first assessment. Early detection of MCI is important as MCI is seen as a risk factor for AD, particularly when there is a memory impairment either alone (i.e., single-domain amnestic MCI) or in combination with deficits in other domains (i.e., multi-domain amnestic MCI) (Albert et al., 2011; J. Eppig et al., 2020; Manly et al., 2008; Thomas et al., 2020). Individuals diagnosed with MCI are significantly more

likely to progress to AD, and do so at a faster rate than those without MCI (Mitchell & Shiri-Feshki, 2009; Pandya, Clem, Silva, & Woon, 2016). Individuals with MCI who are on the AD trajectory often have AD biomarker levels in between those diagnosed as CU and those with AD (Emily C Edmonds et al., 2015; Olsson et al., 2016). Therefore, if PE-adjustment truly improves diagnostic accuracy (i.e., better separates cases from controls) then PE-adjusted MCI diagnoses should be associated with greater likelihood of conversion to dementia and more AD-like biomarker profiles than diagnoses based on PE-unadjusted scores.

Although PEs are thought to be reduced in those who are MCI at baseline, they may still play a significant role in reversion rates. MCI reversion occur when an individual who is diagnosed with MCI at baseline is diagnosed as CU at follow-up. As PEs increase scores at follow-up, they may be a substantial contributor to reversion rates, but, to my knowledge, they have been only studied peripherally (Malek-Ahmadi, 2016; Thomas et al., 2020).

Those critical of MCI typically cite reversion rates as a key concern. Although 10-12% of individuals with MCI are expected to convert to AD per year, 20-50% of individuals revert from MCI to CU status within 2-5 years (Pandya et al., 2016). Over a similar time frame, an estimated 37-67% of individuals retain their MCI diagnosis (Pandya et al., 2016). These findings have led some to conclude that individuals with MCI are more likely to revert to CU or maintain their MCI status than to convert to dementia each year (Canevelli et al., 2016). Several review articles considering reversion rates in MCI have highlighted the wide range in reversion rates and have suggested that this variability is likely due to multiple factors, including the heterogeneity of MCI criteria and reversible causes such as depression (Canevelli et al., 2016; Malek-Ahmadi, 2016; Pandya et al., 2016). Malek-Ahmadi and Pandya et al. also suggested that reducing reversion rates should be an essential goal of future MCI methodology studies (Malek-Ahmadi,

2016; Pandya et al., 2016). Canevelli et al. and Pandya et al. argued that MCI may be an unstable condition where reversion to normal is expected, and that its use as a prodromal stage of underlying neurodegenerative diseases is questionable (Canevelli et al., 2016; Pandya et al., 2016). Malek-Ahmadi suggested that the utility of MCI diagnosis would benefit from further refinement of statistical methods, the use of sensitive cognitive tests, and greater utilization of biomarkers (Malek-Ahmadi, 2016). I believe that one way of addressing the majority of these author's concerns is improving and implementing methods of PE adjustment.

1.2 Practice Effect Adjustment and Clinical Trials

Nearly all AD clinical trials have focused on treating individuals with dementia in an effort to mitigate or reverse the disease. Unfortunately, the failure rate for these trials is greater than 99% (Anand, Patience, Sharma, & Khurana, 2017; J. L. Cummings, Morstorf, & Zhong, 2014). As a result, there has been a shift toward identifying and targeting individuals at the earliest stages of the disease including at-risk CU and MCI (Alexander, Emerson, & Kesselheim, 2021; Anand et al., 2017; Canevelli et al., 2016; R. Sperling, Mormino, & Johnson, 2014; R. A. Sperling et al., 2014). As noted by Canevelli et al., at least 274 randomized controlled trials were recruiting MCI subjects in 2016 (Canevelli et al., 2016). As such, accurate diagnoses of earlier disease stages are necessary to further the treatment of AD (Edmonds et al., 2018a; J. Eppig et al., 2020; Veitch et al., 2019).

The impact of PEs on early detection of decline and stability of MCI diagnosis has a significant impact on AD clinical trials. A review of PEs in MCI and AD samples noted considerable evidence of PEs (i.e., increased scores) in clinical trials (Goldberg et al., 2015). However, despite recognition that accounting for PEs may potentially improve clinical trials and diagnostic accuracy, there are minimal empirical data on PEs in clinical trials (Calamia et al.,

2012; Duff et al., 2011; Goldberg et al., 2015; Jutten et al., 2020). Moreover, PEs are largely ignored in longitudinal studies, clinical trials, and clinical practice, particularly with respect to diagnosis (Calamia et al., 2012; Goldberg et al., 2015; Heilbronner et al., 2010; Machulda et al., 2017; Mathews et al., 2014). Reviews of MCI reversion rates also consistently note that the instability of MCI diagnoses impairs our ability to treat AD by diluting samples and reducing study power (Canevelli et al., 2016; Malek-Ahmadi, 2016; Pandya et al., 2016).

1.3 Replacement Method of Practice Effect Adjustment

The majority of PE adjustment methods only consider PEs when there is an increase in scores at follow-up. As shown previously (Figures A-C), this assumption does not consider the expected age-related decline in older adult populations. Although review papers have noted that PEs can exist even when there is longitudinal decline in observed performance, as expected within a sample at risk for AD (Salthouse, 2010), few have empirically demonstrated that claim (Goldberg et al., 2015). In such situations, Calamia et al. suggested that the most suitable approach is to utilize replacement participants as it is able to gauge PEs even when performance declines (Calamia et al., 2012; Elman et al., 2018; Rönnlund & Nilsson, 2006). The replacement-participants approach involves recruiting participants for testing at follow-up who are demographically matched to returnees. The only difference between groups is that replacements are taking the tests for the first time whereas returnees are retaking the tests. Comparing scores at follow-up between returnees and replacement participants (with additional adjustment for attrition effects) allows for detection of PEs when observed scores remain stable (Figure B) and even when they decline (Figure C). In both scenarios, scores would have been lower without prior exposure. Thus, the goal is to create follow-up scores over retest intervals that are free of PEs and comparable to general normative data. Using a replacement-participants method, in

what to my knowledge is the only study using PE adjustment to modify diagnosis, Elman et al showed that MCI incidence doubled (4.6% v s 9.0%) when scores were adjusted for PEs in a 6-year follow-up study (Elman et al., 2018). The increased incidence means earlier detection of MCI, an important strength of this method.

Despite the utility of the participant replacement method, there are several concerns limiting its use. First, this method lowers all scores in the sample. Therefore, when comparing PE-unadjusted MCI rates to those after PE-adjustment, it is impossible for there to be *decline* in MCI diagnoses at follow-up. It is crucial to determine if the PE-adjusted increased MCI incidence truly represents more accurate diagnosis rather than methodological artifact. Second, although the Elman et al study found an increased rate of MCI after adjustment, there were no follow-up data to confirm that these participants were on a dementia trajectory (Elman et al., 2018). Third, this study did not provide biomarker data to support diagnoses. Fourth, this method requires the recruitment of specifically matched participants at each time point. For many ongoing or completed studies, this may not be feasible. Fifth, to our knowledge, the replacement-participant approach has only been utilized across two timepoints (Elman et al., 2018; Ronnlund, Nyberg, Backman, & Nilsson, 2005). As normal and abnormal aging is inherently a longitudinal process, the viability of a PE-method that is only useable across 2 timepoints is somewhat restricted.

2. Proposed Project

The proposed project represents 3 separate papers that will be combined into a single dissertation that adapts and validates a method of PE adjustment based on the replacement participants method. The first and second papers have been published in peer-reviewed journals. These papers describe the method developed for this dissertation: the pseudo-replacement

method of PE-adjustment. They also use follow-up data and biomarkers to validate how PE adjustment impacts MCI prevalence, reversion, and stability. The third paper will be submitted for peer review with the approval of this dissertation committee. The third paper expands the pseudo-replacement method for use across 3 timepoints to delineate PEs from true cognitive decline by integrating generalized estimating equation (GEE) models. Links to the first and second paper are provided below. A draft of the third paper has been provided below. An integrative discussion follows to comment on conclusions, limitations, related projects, and future directions.

Chapter 1: Cognitive practice effects delay diagnosis of MCI: implications for clinical trials

Authors: Mark Sanderson-Cimino, M.S.,[a,b] Jeremy A. Elman, PhD.,[b,c] Xin M. Tu, PhD.,[c,d,i] Alden L. Gross, PhD.,[e] Matthew S. Panizzon, PhD.,[b,c] Daniel E. Gustavson, PhD.,[f] Mark W. Bondi, PhD.,[c,g] Emily C. Edmonds, PhD.,[c,h] Graham M.L. Eglit, PhD.,[b,c,i] Joel S. Eppig, PhD.,[j] Carol E. Franz, PhD.,[b,c] Amy J. Jak, PhD.,[b,k] Michael J. Lyons, PhD.,[l] Kelsey R. Thomas, PhD.,[c,h] McKenna E. Williams, M.A.,[a,b], William S. Kremen, PhD.,[b,c] , Alzheimer's Disease Neuroimaging Initiative

[a] San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA, USA

[b] Center for Behavior Genetics of Aging, University of California, San Diego, La Jolla, CA, USA

[c] Department of Psychiatry, School of Medicine, University of California, San Diego, La Jolla, CA, USA

[d] Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA

[e] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MA, USA

[f] Department of Medicine, Vanderbilt University Medical Center, Nashville TN, USA

[g] Psychology Service, VA San Diego Healthcare System, San Diego, CA, USA

[h] Research Service, VA San Diego Healthcare System, San Diego, CA, USA

[i] Sam and Rose Stein Institute for Research on Aging, University of California San Diego, La Jolla, CA, USA

[j] VA Puget Sound, Seattle Division, WA, USA

[k] Center of Excellence for Stress and Mental Health, Veterans Affairs San Diego Healthcare System, La Jolla, CA, USA

[l] Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA

Corresponding author: William S. Kremen, Ph.D. Phone: 858-822-2393; Email: wkremen@health.ucsd.edu; Address: Department of Psychiatry (MC 0738), University of California, San Diego, La Jolla, CA 92093

Declaration of interests: Dr. Bondi receives royalties from Oxford University Press, consulting

fees from Roche and Novartis, and honorarium from the International Neuropsychological Society. He is also associated with a clinical trial at the Cleveland Clinic and is the vice chair of the Cognition Professional Interest Area (iSTAART/Alzheimer's Association). Dr. Gross is on the Data Safety Monitory Board for a NIA-funded trial let by Pennington Labs. Dr. Jak receives payments from the Ohio Council on Aging. All other authors declare no competing interests.

Abstract

INTRODUCTION: Practice effects (PEs) on cognitive tests obscure decline, thereby delaying detection of mild cognitive impairment (MCI). Importantly, PEs may be present even when there are performance declines, if scores would have been even lower without prior test exposure. We assessed how accounting for PEs using a replacement-participants method impacts incident MCI diagnosis.

METHODS: Of 889 baseline cognitively normal (CN) Alzheimer's Disease Neuroimaging Initiative (ADNI) participants, 722 returned 1 year later (mean age=74.9±6.8 at baseline). The scores of test-naïve demographically-matched "replacement" participants who took tests for the first time were compared with returnee scores at follow-up. PEs—calculated as the difference between returnee follow-up scores and replacement participants scores—were subtracted from follow-up scores of returnees. PE-adjusted cognitive scores were then used to determine if individuals were below the impairment threshold for MCI. CSF amyloid-beta, phosphorylated-tau, and total tau were used for criterion validation. In addition, based on screening and recruitment numbers from a clinical trial of amyloid-positive individuals, we estimated the effect of earlier detection of MCI by accounting for cognitive PEs on a hypothetical clinical trial in which the key outcome was progression to MCI.

RESULTS: In the ADNI sample, PE-adjusted scores increased MCI incidence by 19% ($p<.001$), increased proportions of amyloid-positive MCI cases (+12%), and reduced proportions of amyloid-positive CNs (-5%) ($p's<.04$). Additional calculations showed that the earlier detection and increased MCI incidence would also substantially reduce necessary sample size and study duration for a clinical trial of progression to MCI. Cost savings were estimated at approximately $5.41 million.

DISCUSSION: Detecting MCI as early as possible is of obvious importance. Accounting for cognitive PEs with the replacement-participants method leads to earlier detection of MCI, improved diagnostic accuracy, and can lead to multi-million-dollar cost reductions for clinical trials.

Keywords: Early diagnosis; Mild cognitive impairment; Practice effects; Alzheimer's disease; Clinical trials; Longitudinal aging

1. INTRODUCTION

Alzheimer's Disease (AD) is a leading cause of death in adults over age 65 and an estimated 1 in 85 people living with the disease by 2050 ("2018 Alzheimer's disease facts and figures," 2018; Brookmeyer, Johnson, Ziegler-Graham, & Arrighi, 2007). Give the protracted AD prodromal period, emphasis is now on clinical trials that begin with cognitively normal (CN) individuals who may progress to mild cognitive impairment (MCI) (J. Cummings, Lee, Ritter, Sabbagh, & Zhong, 2019; Gauthier et al., 2016; Rafii & Aisen, 2019; R. A. Sperling et al., 2014). Delayed detection of MCI is essentially misdiagnosis, i.e., labeling someone as CN when they, in fact, have MCI. Such misdiagnosis impedes identification of meaningful drug effects and may lead to misinterpretation of findings in clinical trials (Bondi et al., 2014; Edmonds et al., 2018b). Clinically, any effects to slow disease progression require early detection. Detection of MCI as early as possible is thus critical.

Repeat cognitive assessments are necessary for accurately determining transitions from CN to MCI or MCI to dementia. However, repeat assessments are subject to practice effects (PEs) that can inflate follow-up scores via memory of specific stimuli (i.e., content PE) or through increased comfort with test taking (i.e., context PE) (Gross et al., 2017; Heilbronner et al., 2010). Put simply, someone taking a test for the second time will typically have a higher score than if they were taking it for the first time. PEs have a wide-ranging impact on any study or field involving cognitive testing because they mask true cognitive decline and compromise diagnostic accuracy, impairing the separation of cases (i.e., MCI) and controls (i.e., CN) (Calamia et al., 2012; Elman et al., 2018). Moreover, PEs are pervasive; they have been found across multiple cognitive domains and test-retest intervals as long as 7 years in older adults, including those with MCI and mild AD (Elman et al., 2018; Goldberg et al., 2015; Gross et al.,

2017; Gross et al., 2015) .

A major limitation of most PE methods is that they only consider PEs when scores are higher at follow-up than at baseline (Calamia et al., 2012; Duff et al., 2011; Goldberg et al., 2015). However, PEs can exist when there is no overall change and when there is decline, as they may still cause underestimation of decline (Figure 1) (Calamia et al., 2012; Goldberg et al., 2015). In such situations, failure to account for PEs may delay MCI diagnosis because PEs would inflate scores above diagnostic impairment thresholds (Calamia et al., 2012; Elman et al., 2018; Rönnlund & Nilsson, 2006; Ronnlund et al., 2005). This is particularly relevant for older adults for whom decline over time may be the norm.

Despite their importance, PEs are largely ignored in longitudinal studies, clinical trials, and clinical practice, particularly with respect to diagnosis (Calamia et al., 2012; Goldberg et al., 2015; Heilbronner et al., 2010; Machulda et al., 2017; Mathews et al., 2014). A review of PEs in MCI and AD samples noted considerable evidence of PEs (i.e., increased scores) in clinical trials.(Goldberg et al., 2015) However, despite recognition that accounting for PEs may potentially improve clinical trials and diagnostic accuracy, there are minimal empirical data on PEs in clinical trials (Calamia et al., 2012; Duff et al., 2011; Goldberg et al., 2015; Jutten et al., 2020).

One method of PE adjustment, the replacement-participant method, is able to gauge PEs even when performance declines (Elman et al., 2018; Rönnlund & Nilsson, 2006). This method relies on the recruitment of an additional set of test-naïve participants (i.e., “replacements”) that is similar to returnees in terms of age and other demographic factors. A comparison of replacement’s performance and that of the returnees calculates a PE because score differences are due to the fact that returnees have taken the tests twice, but replacements have taken the tests

only once. PE-adjusted scores can then be derived by subtracting PEs from the returnee's follow-up scores (Elman et al., 2018; Rönnlund & Nilsson, 2006). Using a replacement-participants method, in what to our knowledge is the only study using PE adjustment to modify diagnosis, we showed that MCI incidence doubled (4.6% v s 9.0%) when scores were adjusted for PEs in a 6-year follow-up study (Elman et al., 2018). The increased incidence means earlier detection of MCI, suggesting an important strength of this method. However, as this method lowers all scores, it is crucial to determine if the increased incidence truly represents more accurate diagnosis rather than methodological artifact. We propose improved correspondence between AD biomarkers and MCI diagnoses as a way of validating the PE-adjusted diagnoses. Other strengths of the method are that returnees and replacements are always well-matched for demographics, and PEs are always calculated based on the specific time interval and for a specific test. A shortcoming of this method is that each test's PE is the same for all subjects because they are group mean effects.

Here, we employed a novel approach by identifying the equivalent of replacement participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI). In individuals who were CN at baseline, we hypothesized that: 1) we would observe PEs at the 12-month follow-up; and 2) accounting for PEs would increase the number of MCI diagnoses at follow-up. Regarding criterion validity, we hypothesized that: 3) PE-adjusted diagnoses would result in more AD biomarker-positive MCI cases and fewer biomarker-positive CN individuals than PE-unadjusted diagnoses. Finally, we completed power/sample size calculations, hypothesizing that: 4) accounting for PEs would substantially reduce the number of participants needed for clinical trials. We then applied these estimates to a hypothetical drug trial with progression to MCI as a key outcome using recruitment data from a major clinical trial. Earlier and more accurate

detection should thus have a substantial impact on clinical trials by reducing study duration, attrition, participant and staff burden, and overall cost.

2. MATERIALS AND METHODS

2.1. Participants

Data were obtained from the ADNI database (adni.loni.usc.edu). The ADNI 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 biological markers, and clinical assessment, and neuropsychological measures can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. Participants from the ADNI-1, ADNI-GO, and ADNI-2 cohorts were included. Informed consent was obtained from all participants.

We identified 889 individuals who were CN at baseline; 722 of them returned for a 12-month-follow-up. Mean educational level of returnees was 16 years (SD=2.7), 47% were female, and mean baseline age was 74.9 years (SD=6.8). All participants completed neuropsychological testing at baseline and 12-month follow-up. After accounting for PEs, we re-diagnosed returnees at their 12-month follow-up as CN or MCI.

2.2. Procedures

Six cognitive tests were examined at both baseline and follow-up (mean=12.21 months; SD=0.97): memory (Wechsler Memory Scaled-Revised, Logical Memory Story A delayed recall; Rey Auditory Verbal Learning Test [AVLT] delayed recall); language (Boston Naming Test; Animal Fluency); attention-executive function (Trails A; Trails B). The American National Adult Reading Test provided an estimate of premorbid IQ. Participants completed the same

version of tests at baseline and 12-month visits.

PE-adjusted and PE-unadjusted scores were converted to z-scores based on external norms that accounted for age, sex, and education for all tests except the AVLT (Shirk et al., 2011). Having found no external norms for the AVLT that were appropriate for this sample and accounted for age, sex, and education, the AVLT was z-scored based on ADNI participants who were CN at baseline (n=889). AVLT demographic corrections were based on a regression model that followed the same approach as the other normative adjustments (Shirk et al., 2011).

We focused primarily on MCI diagnosed according to the Jak-Bondi approach, requiring scores on ≥2 tests within the same cognitive domain to be >1 SD below normative means (Bondi et al., 2014; Edmonds et al., 2018b; Jak et al., 2009). To test whether the results were specific to a particular diagnostic approach, we repeated analyses using Petersen MCI criteria (Jak et al., 2009).

Biomarkers included cerebrospinal fluid amyloid-beta (Aβ), phosphorylated tau (p-tau), and total tau (t-tau) collected at baseline. The ADNI biomarker core (University of Pennsylvania) used the fully automated Elecsys immunoassay (Roche Diagnostics). Sample collection and processing have been described previously (Shaw et al., 2009). Cutoffs for biomarker positivity were: Aβ+: Aβ<977 pg/mL; p-tau+: p-tau>21.8 pg/mL; t-tau+: t-tau>270 pg/mL (http://adni.loni.usc.edu/methods) (Elman, Panizzon, Gustavson, Franz, Sanderson-Cimino, Lyons, & Kremen, 2020; Hansson et al., 2018). There were 521 returnees with Aβ, 518 with p-tau, and 519 with t-tau data.

2.3. Practice effect calculation and statistical analysis

Practice effects were calculated using a modified version of a replacement-participants

method.(Elman et al., 2018) Reviews and meta-analysis have noted that almost all studies of PEs considered only observed performance increases (Figure 1A), and recommended replacement-participants methods in situations where decline is expected (Calamia et al., 2012; Goldberg et al., 2015; Rönnlund & Nilsson, 2006). In some situations PEs will not necessarily manifest as improvements for middle-aged and older adults, particularly for individuals on an AD trajectory (Salthouse, 2010). The replacement-participants approach involves recruiting participants for testing at follow-up who are demographically matched to returnees. The only difference between groups is that replacements are taking the tests for the first time whereas returnees are retaking the tests. Comparing scores at follow-up between returnees and replacement participants (with additional adjustment for attrition effects) allows for detection of PEs when observed scores remain stable (Figure 1B) and even when they decline (Figure 1C). In both scenarios, scores would have been lower without prior exposure. Thus, the goal is to create follow-up scores over retest intervals that are free of PEs and comparable to general normative data. By design, this method is equally applicable for any sample and any tests because returnees and replacements are always matched on demographic characteristics, test, and retest interval.

Because ADNI did not have replacements, we used individuals who at baseline were demographically matched to returnees at follow-up. We refer to them as pseudo-replacements. Bootstrapping (5,000 resamples, with replacement) was used to calculate PE values for each cognitive test. Figure 2 demonstrates how participants were matched at each iteration of the bootstrap. Propensity scores (R package: MatchIt) calculated via one-to-one matching were used to identity pseudo-replacements that were similar to returnees, and an additional constraint confirmed that the returnees and pseudo-replacements were matched at a group level (p's >.8) (D. E. Ho, Imai, King, & Stuart, 2011). Practice effects were calculated by comparing the mean

scores of these subsamples at each bootstrapping iteration using equations displayed in Figure 2. The difference score represents the sum of the practice effect and the attrition effect. With actual replacements, the attrition effect accounts for the fact that returnees are often higher-performing or healthier than those who drop out. However, since the pseudo-replacements are similar to returnees, their removal from the baseline sample lowers the mean baseline score among those not chosen to be returnees at that iteration, resulting in an artificially high attrition effect. To ensure a more accurate attrition effects, we calculated the true attrition and retention rates for each test (approximately 16% and 84%, respectively). We then multiplied the mean score of returnees at baseline by the retention rate and the mean score of the remaining baseline participants (i.e., those not chosen as returnees or pseudo-replacements) by the attrition rate. The sum of these values provides a weighted mean for each iteration, which we refer to as the proportional baseline. Finally, the PE for each test equals the difference score minus the attrition effect (Elman et al., 2018; Ronnlund et al., 2005). The PE for each test was then subtracted from each individual's observed (unadjusted) follow-up test score to provide PE-adjusted raw scores.

In summary, this method identifies a comparison sample (pseudo-replacements) who are matched for age and other demographic characteristics to the returnees. The only difference is that returnees have taken the test before and pseudo-replacements have not. Because this analysis uses completed data, creating pseudo-replacements allowed for application of a replacement method of PE-adjustment to an already completed study without requiring new participant recruitment.

Adjusted raw scores at follow-up were converted to z-scores, which were used to determine PE-adjusted diagnoses. In other words, determination of whether an individual was below the impairment threshold was now based on the PE-adjusted scores. McNemar χ^2 tests

were used to compare differences in the proportion of individuals classified as having MCI before versus after adjusting for PEs, and to determine if PE-adjusted diagnoses changed the proportions of biomarker-positive MCI and CN participants. Cohen's d was calculated for each PE by comparing unadjusted and adjusted scores.

To determine the impact of PE adjustment in a clinical trial, we calculated sample size requirements for a hypothetical clinical trial aimed at reducing progression to MCI at 1-year follow-up in amyloid-positive CN individuals using MCI incidence rates from the present study. We performed logistic regressions with drug/placebo as the predictor and diagnosis at follow-up as the outcome. Sample size estimates were determined across a range of drug effects (10%-40% reduction in MCI diagnoses) with α=.05 and power=.80. We then used this information to estimate the effects on required sample size and cost for a variant of the Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4) Study given α=.05, and power=.80. The A4 Study recruited amyloid-positive CN individuals to investigate whether anti-amyloid therapy can delay cognitive decline (R. A. Sperling et al., 2014). Progression to disease is a common and meaningful outcome for clinical trials. For our hypothetical variation of the A4 Study, the outcome of interest was progression to MCI at 1 year rather than just comparing cognitive decline. These analyses were completed within the powerMediation package in R v3.6.1(Qiu & Qiu, 2020; Team, 2019).

3. RESULTS

PE magnitudes varied within and between cognitive domains (Table 1). PE-adjusted scores resulted in 124 (17%) converters to MCI; unadjusted scores resulted in 104 (14%) converters (Table 2A). Thus, there were 19% ($p<.001$) more individuals diagnosed with MCI

after one year when using PE-adjusted scores (Table 2A). Table 2B shows that adjusting for PEs significantly increased the number of biomarker-positive participants who progressed to MCI (+13% to +15%) and decreased the number of biomarker-positive participants who remained CN (-5%). In particular, there was a 12% increase in amyloid-positive MCI cases and a 5% decrease in amyloid-positive CNs. Supplemental tables 1-2 shows results for diagnoses based on Petersen criteria. The pattern is the same as for the Jak-Bondi criteria, and all significant differences remained significant regardless of diagnostic approach.

Next, we showed that the number of participants necessary to determine a significant drug effect was substantially reduced by accounting for PEs across all effect sizes (Figure 3). On average, adjusting for PEs reduced the number of participants required by 15.9% (321 participants) across effect sizes (range=68-1377 participants). The inset within Figure 3 focuses on differences for hypothetical PE-adjusted and unadjusted samples of approximately 1,000.

We then applied our findings to recruitment data from the A4 Study (Figure 4A) (R. A. Sperling et al., 2020). Obtaining the CN, amyloid-positive A4 sample of 1323 required the recruitment of 5.11 times as many people for initial screening (n=6763) and 3.39 times as many people to undergo amyloid PET imaging (4486) (R. A. Sperling et al., 2020). Our calculations showed that this sample size of 1323 would be powered to detect a 24.7% drug effect on incident MCI outcomes, but accounting for PEs would yield the same power with only 1116. As shown in Figure 4A, the number of initial screens and amyloid PET scans would, in turn, be substantially reduced to 5704 and 3784, respectively. Figure 4B shows the range of sample size reductions for differing drug effect sizes for initial screening (reduced ns range from 347 to 7039) and amyloid PET imaging (reduced ns range from 230 to 4670). As estimated drug effect sizes become smaller, the reductions in necessary sample size become substantially larger, which should lead

to substantial cost reductions.

4. DISCUSSION

Delayed detection of MCI is extremely costly from a public health perspective. In 2018, the Alzheimer's Association projected an estimated U.S. national savings of $231 billion by 2050 if those on the AD trajectory were diagnosed during the MCI, rather than the dementia, stage ("2018 Alzheimer's disease facts and figures," 2018). In clinical practice, the MCI stage represents a critical time for preparation and intervention for individuals who will progress to AD-related dementia. If PEs delay detection of MCI, clinicians may also be providing inadequate care to those most at risk.

Results of the present study confirm our hypothesis that adjusting for PEs using the replacement-participants method does lead to earlier detection of MCI. Accounting for cognitive PEs resulted in a 26% increase in 12-month MCI incidence. The increase in biomarker-positive MCI (+20% amyloid-positive) and reduction in biomarker-positive CN participants (-6% amyloid-positive) supports diagnostic validity. Failure to account for PEs led to a substantial number of false negatives as 18% of biomarker-positive MCI cases were labeled as CN at follow-up. Accounting for PEs improved accuracy, reducing false positives by 5%. Individuals diagnosed with MCI based on PE-adjusted scores—who would otherwise have been considered CN—would be expected to progress to AD dementia sooner than true CN participants. Progression at later follow-ups was consistent with this hypothesis, but sample sizes were too small for statistical comparisons (see supplemental Table 3). Taken together, these results demonstrate that this approach reduces the observed discrepancy between biologically- and clinically-based diagnoses (C. R. Jack et al., 2019). As such, not adjusting for PEs weakens our ability to accurately determine the effect of novel treatments and to compare case-control

biomarker differences, a goal of current research guidelines (Jack Jr et al., 2018). Importantly, the replacement-participant method is not dependent on diagnostic approach. All significant differences for Jak-Bondi criteria remained for Petersen criteria.

To quantify how clinical trials would be improved by PE adjustment, we estimated sample sizes necessary to power a simulated clinical drug trial. Our PE adjustment increased the base rate of MCI at 12-month follow-up and, other things being equal, detecting differences or making predictions is less accurate for low base rate events (Mehta et al., 2009). Progression to disease is the most common outcome of interest in clinical trials, and smaller samples would be needed for clinical trials with a PE-adjusted diagnostic endpoint. Across effect sizes, there was an average reduction of 16% in necessary sample size when using PE-adjusted diagnoses; sample size reductions were greater with smaller treatment effect sizes (Figures 3, 4a). Based on screening/recruitment numbers in the A4 Study (R. A. Sperling et al., 2020). Figure 4A showed that determining progression to MCI using PE-adjusted scores would mean 1060 fewer initial screenings and 703 fewer amyloid PET scans. At $5000 per scan, cost savings for that alone would be $3.52 million. Initial screening—which included cognitive testing, clinical assessments, and *APOE* genotyping—for 1060 individuals would result in considerable additional cost savings, estimated at $2.50 million. Cost saving would be partially offset by needing to add replacement participants. In prior work, 150-200 replacement participants was sufficient (Elman et al., 2018). With replacements for 3 follow-up cognitive assessment sessions with 200 participants each, we estimated additional costs of $615,000. Estimated overall savings would be $5.41 million. Moreover, PE-adjusted diagnoses result in earlier detection, which mean shorter follow-up periods. Reduced study duration would lead to still further cost reductions and benefits including lower participant and staff burden, fewer invasive procedures, and likely

reduced attrition.

The present study may raise the question of how the replacement-participants method compares with other approaches to PEs, but that is likely to be the wrong question because different approaches may be for different purposes. A 2012 meta-analysis and 2015 review described several approaches to estimating PEs (Calamia et al., 2012; Goldberg et al., 2015). Almost all non-replacement approaches—including more commonly employed regression-based approaches—are only informative about *relative* differences, including predicting future change. One interesting paradigm is to retest participants after a short interval using a regression-based approach, and then have much a longer follow-up. Individuals with smaller PEs at 1 week are more likely to have worse baseline biomarker profiles, experience steeper 1-year decline, and progress to MCI/AD compared to other participants (Duff, 2014; Duff, Foster, & Hoffman, 2014; Duff et al., 2011; Jutten et al., 2020). Thus, this approach may be useful for participant selection in clinical trials. Other studies have found that additional baseline tests improve prediction of progression to MCI (Bondi et al., 2014; Emily C Edmonds et al., 2015; Edmonds et al., 2016; Elman, Vuoksimaa, Franz, & Kremen, 2020; Gustavson et al., 2020; Jutten et al., 2020; Vuoksimaa, McEvoy, Holland, Franz, & Kremen, 2020). Whether complete 1-week retesting of the entire sample improves prediction over the less burdensome and less costly inclusion of additional measures at baseline testing remains to be determined. Also, regression-based methods require a large, normative change sample, and would require new, large normative samples for each study if the specific tests, retest intervals, or sample demographics are different. Change is assessed relative to the normative sample, but PEs are still unknown in the normative sample. Importantly, regression-based approaches cannot be used for absolute diagnostic cutoff thresholds because, unlike the replacement-participant method, they do not

produce stand-alone follow-up scores adjusted for PEs. Thus, they cannot have any effect on *when* a person crosses an impairment threshold and cannot lead to earlier detection of conversion to MCI. Nor can they calculate PEs in the presence of a mean-level decline over time, which is expected in older adults. The replacement-participants method requires a small number of additional participants relative to an entire study sample, and it generates adjusted scores at follow-up that are not obscured by age-related decline. The other methods can compare trajectories of people already diagnosed as MCI or CN, but only the replacement method—which generates absolute PE-adjusted scores—can alter when MCI is detected. Although the replacement-participants method reduces all scores, it does not change individual differences in any way. Thus, it also allows for comparison of trajectories. More thorough discussions of PE methods can be found in a systematic review by Calamia[11], the position paper on PEs by the American Academy of Clinical Neuropsychology[10], and a study that directly compares three regression-based PE approaches[42].

We acknowledge some limitations of the study. ADNI is not a population-based study and is not representative of the general population in terms of sociodemographic factors. However, replacement methods have been shown to be effective in other studies, including population-based samples (Elman et al., 2018; Ronnlund et al., 2005). The method currently only examines PEs across 2 time points. As PEs persist over time, their magnitude may differ with additional assessments. Future studies should explore PEs in cases with multiple follow-up visits. As noted, including matched replacements for third and fourth visits would still be cost-effective. Some participants who do not qualify after initial screening or those who do not agree to biomarker assessment might still qualify to serve as replacement participants. Importantly, the PE magnitudes in the present study should not be directly used in other studies. PEs are often

sample specific and need to be calculated with appropriate replacement participants for each study (Calamia et al., 2012; Duff & Hammers, 2020). Ultimately, the field may benefit from the development of PE norms for standard neuropsychological tests at some clinically meaningful intervals (e.g., 6 and 12 months). This could reduce the need for replacement participants. Similarly, sample size estimations for our hypothetical clinical trial may not be the same for other studies, but do provide more empirical evidence supporting the use of PE-adjustment.

Surprisingly, we found no practice effect on the AVLT. This may have occurred because, despite receiving the same version at the 12-month visit, some participants also completed an alternate version of the AVLT at a 6-month visit. The reduced 12-month practice effect for AVLT is consistent with the well-known phenomenon of retroactive interference, i.e., the different 6-month version interfering with the PE from exposure to the baseline/follow-up version. Prior studies, including our own, have consistently found PEs on the AVLT or similar episodic memory measures (Calamia et al., 2012; Elman et al., 2018; Ronnlund et al., 2005). Thus, the present estimate of the impact of PEs may be a conservative one. It is also noteworthy that despite the lack of an apparent AVLT practice effect in the current study, we still found an increase in amnestic MCI cases after adjusting for PEs. This highlights the importance of including more than one test in each cognitive domain as specified in the Jak-Bondi approach (Bondi et al., 2014; Edmonds et al., 2018b; Elman et al., 2018; Gustavson et al., 2020). Finally, we note that use of alternate forms is considered suboptimal as even well-matched forms are not equivalent and add an additional source of test-retest variance (Gross et al., 2012).

In summary, adjusting for PEs results in earlier and more accurate detection of MCI. Reluctance to include additional replacement-participant testing is understandable as it increases cost and participant burden. In the end, however, it would substantially reduce the necessary

sample size, follow-up time, participant and staff burden, and cost for clinical trials or other longitudinal studies. Although the magnitude of PEs may not be generalized from one sample to another, the replacement-participant method is appropriate for all ages, tests, and retest intervals because replacements are always matched on these features. The method is also not dependent on any specific approach to the diagnosis of MCI. Additionally, we have shown that the replacement-participant method can be adapted for ongoing or already completed studies that did not recruit matched-replacement participants in advance. Given the public health importance of the earliest possible identification of AD pathology, we strongly recommend that accounting for PEs be a planned component of clinical trials, routine clinical work, and longitudinal studies of aging and aging-related cognitive disorders.

3. ACKNOWLEDGMENTS:

The content of this article is the responsibility of the authors and does not necessarily represent official views of the National institute of Aging or the Department of Veterans affairs. The ADNI and funding sources had no role in data analysis, interpretation, or writing of this project. The corresponding author was granted access to the data by ADNI and conducted the analyses.

Author contributors:

The study was conceived by MSC and WSK. Guidance on statistical analysis was provided by XMT and ALG. Determination of MCI diagnoses was made by ECE, MWB, JE, and KRT. MSC, WSK, JAE, MSP, and DEG contributed to the practice effects methodology. Primary funding to support this work was obtain by WSK, CEF, MJL, and MSC. All authors

provided critical review and commentary on the manuscript.

References

1 2018 Alzheimer's disease facts and figures. (2018). *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 14*(3), 367-429. doi:10.1016/j.jalz.2018.02.001

2 Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., . . . Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *J Alzheimers Dis*. doi:10.3233/JAD-140276

3 Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia, 3*(3), 186-191.

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.

5 Cummings, J., Lee, G., Ritter, A., Sabbagh, M., & Zhong, K. (2019). Alzheimer's disease drug development pipeline: 2019. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 5*, 272-293.

6 Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical neuropsychologist, 28*(5), 714-725.

7 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.

8 Duff, K., & Hammers, D. B. (2020). Practice effects in mild cognitive impairment: A validation of Calamia et al.(2012). *The Clinical neuropsychologist*, 1-13.

9 Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., . . . McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry, 19*(11), 932-939.

10 Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2018). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: A secondary analysis of the ADCS vitamin E and donepezil in MCI study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*, 11-18.

11 Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., . . . Salmon, D. P. (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia, 11*(4), 415-424.

12 Edmonds, E. C., Delano-Wood, L., Jak, A. J., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2016). "Missed" mild cognitive impairment: High false-negative error rate based on conventional diagnostic criteria. *Journal of Alzheimer's Disease, 52*(2), 685-691.

13 Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., . . . Jacobson, K. C. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*.

14 Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., & Kremen, W. S. (2020). Amyloid-β Positivity Predicts Cognitive Decline but Cognition Predicts Progression to Amyloid-β Positivity. *Biological Psychiatry*.

15 Elman, J. A., Vuoksimaa, E., Franz, C. E., & Kremen, W. S. (2020). Degree of cognitive impairment does not signify early versus late mild cognitive impairment: confirmation based on Alzheimer's disease polygenic risk. *Neurobiology of aging, 94*, 149-153.

16 Gauthier, S., Albert, M., Fox, N., Goedert, M., Kivipelto, M., Mestre-Ferrandiz, J., & Middleton, L. T. (2016). Why has therapy development for dementia failed in the last two decades? *Alzheimer's & Dementia, 12*(1), 60-64.

17 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.

18 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: The Journal of the Alzheimer's Association, 13*(7), P473-P474.

19 Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., . . . 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.

20 Gross, A. L., Inouye, S. K., Rebok, G. W., Brandt, J., Crane, P. K., Parisi, J. M., . . . Jones, R. N. (2012). Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. *Journal of Clinical and Experimental Neuropsychology, 34*(7), 758-772.

21 Gustavson, D. E., Elman, J. A., Sanderson-Cimino, M., Franz, C. E., Panizzon, M. S., Jak, A. J., . . . Kremen, W. S. (2020). Extensive memory testing improves prediction of progression to MCI in late middle age. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1).

22 Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J. Q., Bittner, T., . . . Alzheimer's Disease Neuroimaging, I. (2018). CSF biomarkers of Alzheimer's disease concord with amyloid-beta PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. doi:10.1016/j.jalz.2018.01.010

23 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.

24 Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw, 42*(8), 1-28.

25 Jack, C. R., Therneau, T. M., Weigand, S. D., Wiste, H. J., Knopman, D. S., Vemuri, P., . . . Machulda, M. M. (2019). Prevalence of biologically vs clinically defined Alzheimer spectrum entities using the National Institute on Aging–Alzheimer's Association research framework. *JAMA neurology, 76*(10), 1174-1183.

26 Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Karlawish, J. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562.

27 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 : official journal of the American Association for Geriatric Psychiatry, 17*(5), 368-375. doi:10.1097/jgp.0b013e31819431d5

28 Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A., Jones, R. N., Choi, S. E., . . . Tommet, D. (2020). Lower practice effects as a marker of cognitive performance and

dementia risk: a literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1), e12055.

29 Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., . . . 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.

30 Mathews, M., Abner, E., Kryscio, R., Jicha, G., Cooper, G., Smith, C., . . . 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.

31 Mehta, C., Gao, P., Bhatt, D. L., Harrington, R. A., Skerjanec, S., & Ware, J. H. (2009). Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation, 119*(4), 597-605.

32 Qiu, W., & Qiu, M. W. (2020). Package 'powerMediation'.

33 Rafii, M. S., & Aisen, P. S. (2019). Alzheimer's Disease Clinical Trials: Moving Toward Successful Prevention. *CNS drugs, 33*(2), 99-106.

34 Rönnlund, M., & Nilsson, L.-G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence, 34*(1), 63-78.

35 Ronnlund, M., Nyberg, L., Backman, 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. *Psychol Aging, 20*(1), 3-18. doi:10.1037/0882-7974.20.1.3

36 Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society, 16*(5), 754-760.

37 Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., . . . Alzheimer's Disease Neuroimaging, I. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol, 65*(4), 403-413. doi:10.1002/ana.21610

38 Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy, 3*(6), 32.

39 Sperling, R. A., Donohue, M. C., Raman, R., Sun, C.-K., Yaari, R., Holdridge, K., . . . Aisen, P. S. (2020). Association of factors with elevated amyloid burden in clinically normal older individuals. *JAMA neurology*.

40 Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: stopping AD before symptoms begin? *Science translational medicine, 6*(228), 228fs213-228fs213.

41 Team, R. C. (2019). language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

42 Vuoksimaa, E., McEvoy, L. K., Holland, D., Franz, C. E., & Kremen, W. S. (2020). Modifying the minimum criteria for diagnosing amnestic MCI to improve prediction of brain atrophy and progression to Alzheimer's disease. *Brain imaging and behavior, 14*(3), 787-796.

Figures and tables

Table 1: Means, standard deviations, attrition effects, and practice effects for each cognitive test.

	Memory		Attention/Executive Function		Language	
Raw mean score (SD)	**RAVLT**	**Logical Memory**	**Trails A**	**Trails B**	**Boston Naming**	**Category Fluency**
Proportional Baseline	7.18 (3.81)	10.64 (4.24)	31.89 (10.79)	77.47 (39.86)	29.04 (2.42)	19.67 (5.23)
Returnees Baseline	7.18 (3.79)	10.54 (4.23)	31.97 (10.82)	76.89 (38.41)	29.05 (2.36)	19.71 (5.26)
Returnees Follow-Up	6.97 (4.38)	11.66 (4.63)	31.52 (12.52)	75.24 (43.14)	29.43 (2.27)	19.84 (5.22)
Replacements Follow-Up	6.97 (3.79)	10.60 (4.34)	32.47 (10.83)	79.38 (41.69)	28.99 (2.45)	19.46 (5.19)
Attrition Effect	0	-.09	-.02	-.59	.01	.03
Practice Effect	NA	1.15	-.93	-3.56	.43	.35
Cohen's d	NA	.24	-.07	-.08	.19	.07

Groups are based on the average performance across all 5000 bootstrapped iterations. Means are based on transformed data that was reverted back to raw units. "Proportional baseline" refers to a weighted mean that combines the returnee baseline group and a group that included all subjects not selected to be Returnees or Replacements in that bootstrapped iteration. "Returnee Baseline" refers to baseline test scores for the portion of participants who returned for the 12-month follow-up visit (n=722). "Returnee Follow-Up" refers to 12-month scores for the portion of participants who returned for the 12-month follow-up (n=722). "Replacement Follow-up" refers to the pseudo-replacement scores. The scores for memory tasks indicate the number of words remembered at the delayed recall trials. Scores on the attention/executive functioning tests indicate time to completion of task. On these tasks, higher scores indicate worse performance. Scores on the Boston Naming Task indicate number of correct items identified; scores on Category Fluency indicate number of items correctly stated. Practice effects and attrition effects are in raw units. As such, the negative practice effects and attrition effects for the Trails tasks demonstrate that practice decreased time (increased performance). Cohen's d is given for the difference between PE-adjusted and unadjusted scores of returnees at follow-up. RAVLT= Rey Auditory Verbal Learning Test.

2A. Progression from cognitively normal to MCI

	# of cases, based on PE-unadjusted cognitive scores	# of cases, based on PE-adjusted cognitive scores	Difference in # of cases (%)	χ^2; *p*
MCI diagnosis	104	124	+20 (+19%)	18.1; <.001
Memory domain impaired	74	87	+13 (+18%)	11.1; <.001
Attention/Executive domain impaired	21	25	+4 (+19%)	2.3; .13
Language domain impaired	11	14	+3 (+27%)	1.3; .25
Impaired on 1 test within all domains	11	13	+2 (+18%)	.17; .68

Follow-up diagnoses were made with practice effect-unadjusted (PE-unadjusted) or practice effect-adjusted (PE-adjusted) scores. The difference in the number of cases is calculated by subtracting the number of cases, based on PE-unadjusted scores, from the number of cases based on PE-adjusted scores. The percent difference (%) in number of cases is the differences in number of cases divided by the number of cases based on PE-unadjusted cognitive scores (e.g., 19%=20/104). χ^2 is McNemar χ^2. Individuals could be impaired in more than one domain. Consequently, the sum of impaired individuals within each domain is greater than the total number of MCI cases. The MCI diagnosis row counts an individual only once, even if they are impaired in more than one domain.

2B. Concordance of MCI diagnosis and biomarker-positivity

Converters to MCI	**# of returnees who are biomarker-positive and MCI (PE-unadjusted)**	**# of returnees who are biomarker-positive and MCI (PE-adjusted)**	**Difference in # of cases (%)**	***p***
Aβ+	51	58	+7 (+14%)	.02
p-tau+	54	62	+8 (+15%)	.01
t-tau+	47	53	+6 (+13%)	.04
Stable CN	**# of returnees who are biomarker-positive and CN (PE-unadjusted)**	**# of returnees who are biomarker-positive and CN (PE-adjusted)**	**Difference in # of cases (%)**	***p***
Aβ+	152	145	-7 (-5%)	.02
p-tau+	170	162	-8 (-5%)	.01
t-tau+	118	112	-6 (-5%)	.04

Follow-up diagnoses were made with practice effect-unadjusted (PE-unadjusted) or practice effect-adjusted (PE-adjusted) scores. The difference in the number of cases is calculated by subtracting the number of cases, based on PE-unadjusted scores, from the number of cases based on PE-adjusted scores. The percent difference (%) in number of cases is the differences in number of cases divided by the number of cases based on PE-unadjusted cognitive scores (e.g., 19%=20/104). χ^2 is McNemar χ^2. Individuals could be impaired in more than one domain. Consequently, the sum of impaired individuals within each domain is greater than the total number of MCI cases. The MCI diagnosis row counts an individual only once, even if they are impaired in more than one domain.

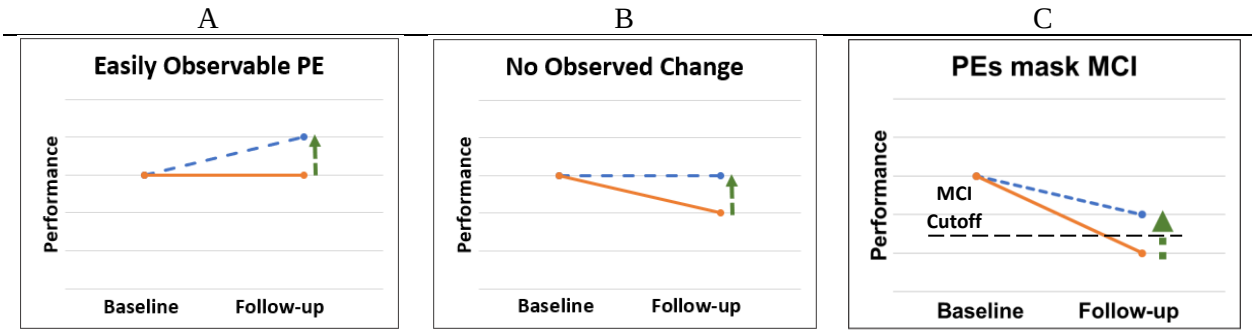


Figure 2: **Practice effects with and without true decline.** The solid line represents true cognitive ability. The dashed line represents observed performance, which is inflated due to a practice effect (vertical arrow). **1A:** Typically observed practice effect: an individual's observed score increases from baseline to follow-up, demonstrating a typical practice effect. **1B**: Practice effect in the context of cognitive decline. In this scenario, an individual's ability is decreasing overtime. A practice effect still exists but is masked by cognitive decline. As a result, the individual's performance appears to be stable but is actually better than it would have been without previous exposure to the test. **1C**: Practice effects impair detection of MCI. In this situation, an individual has declined below an MCI cutoff. However, practice effects are inflating their score so that they now fall above the MCI cutoff and will be diagnosed as cognitively normal at follow-up.

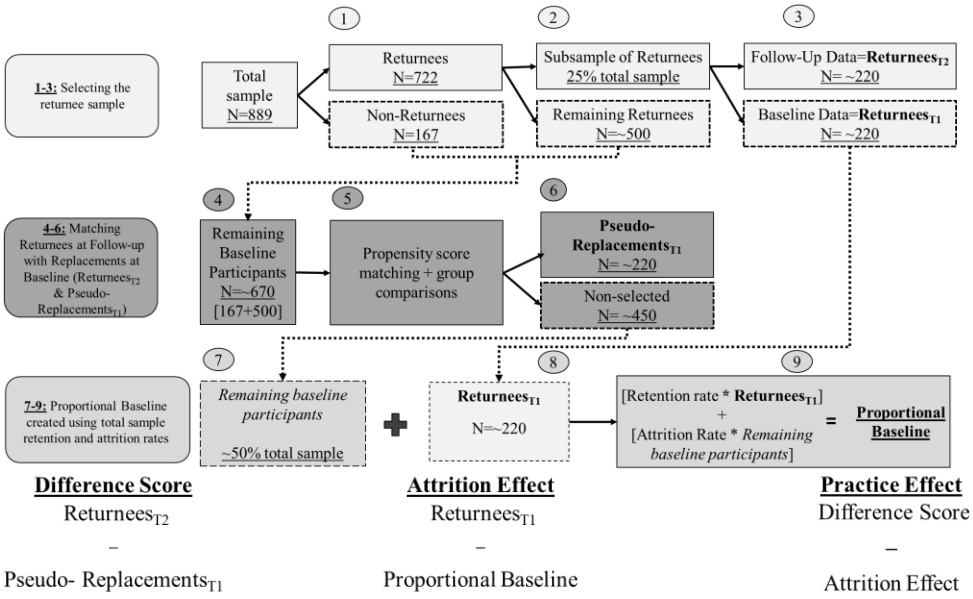


Figure 3: Sample matching and practice effect calculations. Practice effect calculations are based on bootstrapped analyses. Participants with valid baseline data were identified (n=889). **[1]** Participants who also had 12-month follow-up data comprised the returnees (n=722). **[2]** A subsample (n=25% total sample) of returnees was selected; this was approximately 220 participants. **[3]** Baseline data for these participants were labeled as **Returnees$_{T1}$**. Follow-up data for these participants were labeled **Returnees$_{T2}$. [4]** The 220 Returnees$_{T1}$ participants were removed from the pool of baseline data, leaving approximately 670 remaining baseline participants. **[5]** Using propensity score matching with an additional age restriction (<0.1 years), the potential pseudo-replacements were matched to the Resturnes$_{T2}$ participants using one-to-one matching. The pseudo-replacements were drawn from the 670 remain baseline participant pool. Matching parameters were age, birth sex, education, and premorbid IQ. Additionally, comparisons of age, birth sex, education, and premorbid IQ were to confirm groups were similar (p's >.80). **[6]** Once matching was complete, the sample was labeled **Pseudo-Replacments$_{T1}$**, and this sample ranged in size from 200-240 participants. Thus, the Pseudo-Replacmensts$_{T1}$ sample and the Returnees$_{T2}$ sample were demographically matched and only differed in that the Returnees$_{T2}$ had taken the test before while Pseudo-Replacmeents$_{T1}$ had taken the tests only once. After the—on average—220 Pseudo-Replacments$_{T1}$ were removed from the pool of baseline data, there were 450 remaining unchosen baseline participants, or 50% of the total sample. The previous steps were completed at each of the 5000 iterations.

Practice effects were calculated by comparing the mean scores of these subsamples using the equations provided below the flowchart. The difference between the mean of Returnees$_{T2}$ scores and the mean of the matched Pseudo-Replacements$_{T1}$ scores equates to the sum of practice effect and attrition effect. The attrition effect accounts for the fact that individuals who return for follow-up may be a higher performing or healthier than the full baseline sample. **[7-9]** To retain the proportion of returnees to attritors we had in the original sample, we then created a weighted mean of the baseline data cognitive score by multiplying the mean test score of the remaining baseline subject pool by the attrition rate (approximately 16%) and the Returnees$_{T1}$ pool by the retention rate (84%); this is referred to as the **Proportional Baseline** in the text. The practice effect for each test equals the difference score minus the attrition effect.

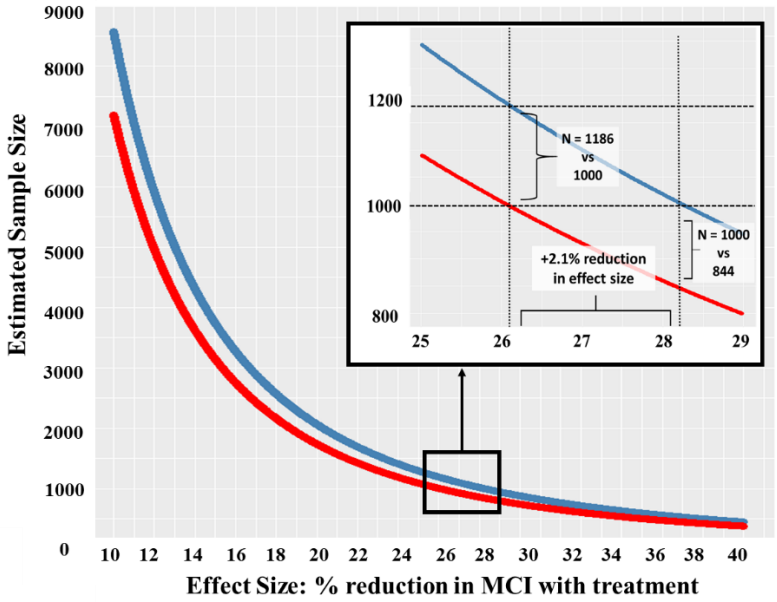


Figure 4: Effect of practice effect-adjusted vs unadjusted scores on a hypothetical clinical trial of biomarker-positive participants. Comparison of estimated sample sizes (Y-axis) necessary for detecting a significant drug effect (X-axis) in a sample that is biomarker-negative and cognitively normal at baseline. The drug effect is operationalized as percent reduction in mild cognitive impairment (MCI) diagnoses at a 1-year follow-up between the treatment group and the placebo group. For example, a drug effect of 30% means that 30% more participants remained cognitively normal when treated with the drug than when given the placebo.
The red line represents a trial that uses MCI incidence rates based on practice effect (PE)-adjusted diagnoses and the blue line represents a trial that uses incidence rates based on unadjusted diagnoses. MCI incidence rates were based on the subsample of participants from the present study who were biomarker-negative and cognitively normal at baseline. The model examined was a logistic regression with diagnosis at follow-up (MCI vs cognitively normal) as the outcome variable. The predictor was a two-level categorical variable representing placebo or drug. Alpha was set at .05, power was .80, and the hypothetical sample was evenly split into treatment and placebo groups.
Across all effect sizes (10%-40% reduction in treatment vs placebo conversation rates) the PE-adjusted trial required fewer participants than the PE-unadjusted trial. The inset shows results for hypothetical samples with ~1000 participants. If this study used PE-unadjusted outcome measures (blue line), it would require an effect size of 28.2% to reach a significant result with ~1000 participants. Using PE-adjusted diagnoses, only 844 participants would be required for the same study with the same drug effect, a reduction of 156 participants. A PE-adjusted study with ~1000 participants (red line in the inset) would be able to detect a smaller drug effect of 26.1%. With this 2.1% reduction in effect size, a PE-unadjusted study would require an additional 186 participants at this drug effect level (1186 vs 1000).

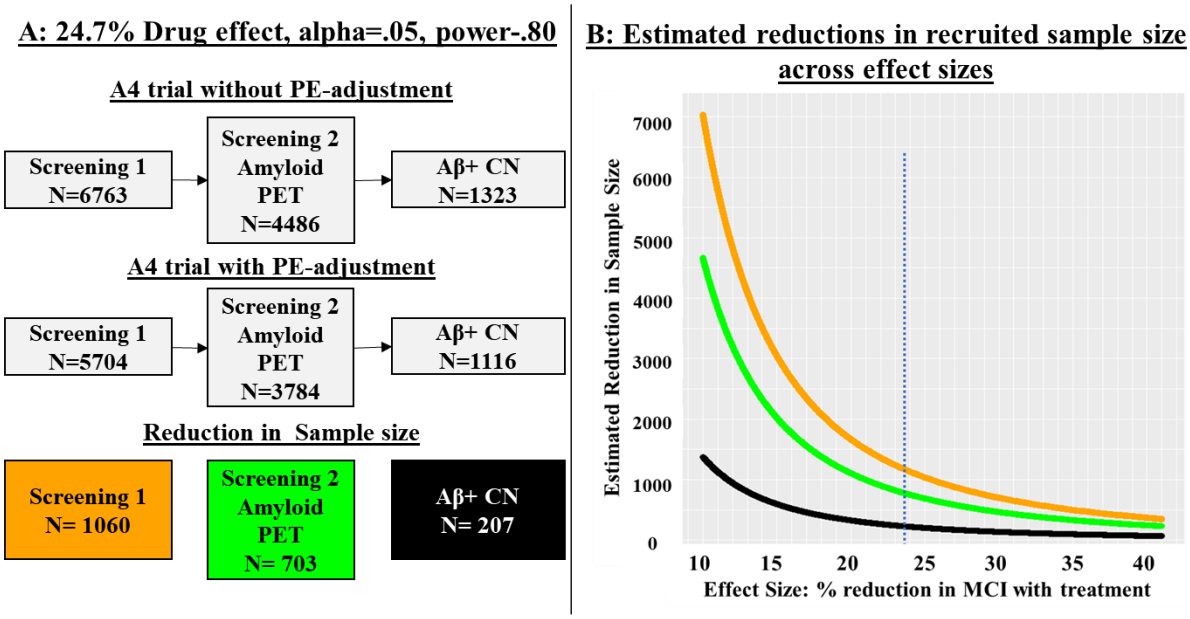


Figure 5: Comparison of recruitment designs for detection of a drug effect based on A4 Study recruitment. Using sample size estimates from Figure 3, we present how planning to adjust for practice effects would alter a clinical drug trial, using A4 Study recruitment as an example. The A4 Study had a total sample of 1323 participants after recruitment as shown in the top row of gray boxes (based on Figure 1 in Sperling et al., 2020).[30] **A**: Based on sample size estimates from Figure 3, a sample of 1323 would enable a study to detect a significant drug effect of 24.7% at an alpha of .05 and .80 power. The top row of the flow chart presents the recruitment for the A4 study. This study reported an initial screening (6763 participants) followed by amyloid PET (4486 participants) imaging to achieve their sample of 1323 amyloid-positive (AB+), cognitively normal (CN) participants. Achieving the final sample size thus required an n for the initial screening that was 5.11 times as large as the final sample size, and an n for amyloid PET imaging that was 3.39 times as large as the final sample. Our power analyses suggest that the same effect size is achieved with only 1116 participants if a trial adjusted follow-up scores for practice effects. That, along with the reductions in initial screening and PET scans, is shown in the middle row of the flow chart. The bottom row shows the sample size reductions for initial screening, PET screening, and the initial biomarker-positive and cognitively normal sample. **B**: The figure presents the reduction in recruitment sample size (Y-axis) across effect sizes ranging from 10% to 40% (X-axis). The orange line represents how many fewer participants would be necessary at initial screening if a study had planned to adjust for practice effects at follow-up.

Chapter 1, in full, is a reprint of the material as it appears in Alzheimer's & Dementia: Translational Research & Clinical Interventions. 1. Mark Sanderson-Cimino, Jeremy A. Elman, Xin M. Tu, Alden L. Gross, Matthew S. Panizzon, Daniel E. Gustavson, Mark W. Bondi, Emily C. Edmonds, Graham M.L. Eglit, Joel S. Eppig, Carol E. Franz, Amy J. Jak, Michael J. Lyons, Kelsey R. Thomas, McKenna E. Williams, William S. Kremen. *Cognitive Practice Effects Delay Diagnosis; Implications for Clinical Trials.* Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2021. The dissertation author was the primary investigator and author of this paper.

Chapter 2: Practice effects in mild cognitive impairment increase reversion rates and delay detection of new impairments.

Authors: Mark Sanderson-Cimino, M.S.,[1,2,*] Jeremy A. Elman, PhD.,[2,3] Xin M. Tu, PhD.,[3,4,9] Alden L. Gross, PhD.,[5] Matthew S. Panizzon, PhD.,[2,3] Daniel E. Gustavson, PhD.,[6] Mark W. Bondi, PhD.,[3,7] Emily C. Edmonds, PhD.,[3,8] Joel S. Eppig, PhD.,[10] Carol E. Franz, PhD.,[2,3] Amy J. Jak, PhD.,[2,11] Michael J. Lyons, PhD.,[12] Kelsey R. Thomas, PhD.,[3,8] McKenna E. Williams, M.A.,[1,2], and William S. Kremen, PhD.,[2,3,11]

**for the Alzheimer's Disease Neuroimaging Initiative

[1] San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA, USA

[2] Center for Behavior Genetics of Aging, University of California, San Diego, La Jolla, CA, USA

[3] Department of Psychiatry, School of Medicine, University of California, San Diego, La Jolla, CA, USA

[4] Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA

[5] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MA, USA

[6] Department of Medicine, Vanderbilt University Medical Center, Nashville TN, USA

[7] Psychology Service, VA San Diego Healthcare System, San Diego, CA, USA

[8] Research Service, VA San Diego Healthcare System, San Diego, CA, USA

[9] Sam and Rose Stein Institute for Research on Aging, University of California San Diego, La Jolla, CA, USA

[10] Rehabilitation Institute of Washington, Seattle, WA, USA

[11] Center of Excellence for Stress and Mental Health, Veterans Affairs San Diego Healthcare System, La Jolla, CA, USA

[12] Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA

**Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI

investigators can be found at:

http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

***Corresponding author**: Mark Sanderson-Cimino, M.S. Phone: 925-586-5102; Email: mesander@health.ucsd.edu

Key Words: [5-8]: Practice effects; Cognitive Aging; Mild Cognitive Impairment; Alzheimer's Disease; Biomarkers; Dementia Progression

Abstract

Objective: Cognitive practice effects (PEs) can delay detection of progression from cognitively unimpaired to mild cognitive impairment (MCI). They also reduce diagnostic accuracy as suggested by biomarker positivity data. Even among those who decline, PEs can mask steeper declines by inflating cognitive scores. Within MCI samples, PEs may increase reversion rates and thus impede detection of further impairment. Within an MCI sample at baseline, we evaluated how PEs impact prevalence, reversion rates, and dementia progression after 1 year.

Methods: We examined 329 baseline Alzheimer's Disease Neuroimaging Initiative MCI participants (mean age=73.1; SD=7.4). We identified test-naïve participants who were demographically matched to returnees at their 1-year follow-up. Since the only major difference between groups was that one completed testing once and the other twice, comparison of scores in each group yielded PEs. PEs were subtracted from each test to yield PE-adjusted scores. Biomarkers included cerebrospinal fluid phosphorylated tau and amyloid beta. Cox proportional models predicted time until first dementia diagnosis using PE-unadjusted and PE-adjusted diagnoses.

Results: Accounting for PEs increased MCI prevalence at follow-up by 9.2% (272 vs 249 MCI), and reduced reversion to normal by 28.8% (57 vs 80 reverters). PEs also increased stability of single-domain MCI by 12.0% (164 vs 147). Compared to PE-unadjusted diagnoses, use of PE-adjusted follow-up diagnoses led to a 2-fold increase in hazard ratios for incident dementia. We classified individuals as false reverters if they reverted to cognitively unimpaired status based on PE-unadjusted scores, but remained classified as MCI cases after accounting for

PEs. When amyloid and tau positivity were examined together, 72.2% of these false reverters were positive for at least one biomarker.

Interpretation: Even when PEs are small, they can meaningfully change whether some individuals with MCI retain the diagnosis at a 1-year follow-up. Accounting for PEs resulted in increased MCI prevalence and altered stability/reversion rates. This improved diagnostic accuracy also increased the dementia-predicting ability of MCI diagnoses.

INTRODUCTION

Mild cognitive impairment stability and reversion

Mild cognitive impairment (MCI) is characterized by cognitive deficits in the presence of minimal to no impairment in functional activities (Albert et al., 2011; Manly et al., 2008). MCI is seen as a risk factor for Alzheimer's Disease dementia (AD), particularly when there is a memory impairment either alone (i.e., single-domain amnestic MCI) or in combination with deficits in other domains (i.e., multi-domain amnestic MCI) (Albert et al., 2011; J. Eppig et al., 2020; Manly et al., 2008; Thomas et al., 2020). Individuals diagnosed with MCI are significantly more likely to progress to AD, and do so at a faster rate than those without MCI (Mitchell & Shiri-Feshki, 2009; Pandya et al., 2016). Individuals with MCI who are on the AD trajectory often have AD biomarker levels in between those diagnosed as cognitively normal (CN) and those with AD (Emily C Edmonds et al., 2015; Olsson et al., 2016).

Nearly all AD clinical trials have focused on treating individuals with dementia in an effort to mitigate or reverse the disease. Unfortunately, the failure rate for these trials is greater than 99% (Anand et al., 2017; J. L. Cummings et al., 2014). As a result, there has been a shift toward identifying and targeting individuals at the earliest stages of the disease including at-risk CN and MCI (Alexander et al., 2021; Anand et al., 2017; Canevelli et al., 2016; R. Sperling et al., 2014; R. A. Sperling et al., 2014). As noted by Canevelli et al, at least 274 randomized controlled trials were recruiting MCI subjects in 2016 (Canevelli et al., 2016). As such, accurate diagnoses of earlier disease stages are necessary to further the treatment of AD (Edmonds et al., 2018a; J. Eppig et al., 2020; Veitch et al., 2019).

There is concern regarding stability of MCI diagnosis that limits its use in clinical and research settings. Although 10-12% of those with MCI are expected to convert to AD per year, 20-50% of individuals revert from MCI to CN status within 2-5 years (Pandya et al., 2016). Over a similar time frame, an estimated 37-67% of individuals retain their MCI diagnosis (Pandya et al., 2016). One criticism of the MCI diagnosis has centered on the fact that individuals are more likely to revert to CN or maintain their MCI status than to convert to dementia each year (Canevelli et al., 2016). On the other hand, long term follow-ups may be necessary to more accurately determine the true proportion of those with MCI who progress to dementia.

Much of the MCI reversion rate literature was published prior to 2016 and was summarized by three articles (Canevelli et al., 2016; Malek-Ahmadi, 2016; Pandya et al., 2016). These authors highlighted the wide range in reversion rates and suggested that this variability is likely due to multiple factors, including the heterogeneity of MCI criteria and reversible causes such as depression (Canevelli et al., 2016; Malek-Ahmadi, 2016; Pandya et al., 2016). Malek-Ahmadi and Pandya et al. also suggested that reducing reversion rates should be an essential goal of future MCI methodology studies(Malek-Ahmadi, 2016; Pandya et al., 2016). Canevelli et al and Pandya et al argued that MCI may be an unstable condition where reversion to normal is expected, and that its use as a prodromal stage of underlying neurodegenerative diseases is questionable (Canevelli et al., 2016; Pandya et al., 2016). Malek-Ahmadi suggested that the utility of MCI diagnosis would benefit from further refinement of statistical methods, the use of sensitive cognitive tests, and greater utilization of biomarkers (Malek-Ahmadi, 2016). All three reviews concluded that reversion impairs our ability to treat AD by diluting samples and reducing study power (Canevelli et al., 2016; Malek-Ahmadi, 2016; Pandya et al., 2016).

Practice effects and MCI

Practice effects (PEs) on cognitive tests used to diagnose MCI are a likely contributor to MCI reversion rates. They mask cognitive decline by increasing scores at follow-up testing relative to how an individual would have performed if they were naïve to the test. PEs are due to familiarity with specific test items (i.e., content effect), and/or increased comfort and familiarity with the general assessment process (i.e., context effect) (Calamia et al., 2012; Gross et al., 2017). PEs in participants without dementia have been found across retest intervals as long as 7 years, and across multiple cognitive domains (Elman et al., 2018; Gross et al., 2015; Ronnlund et al., 2005; Wang et al., 2020). PEs after 3-6 months have even been observed in those with mild AD who performed very poorly on memory measures (Goldberg et al., 2015; Gross et al., 2017). Although PEs may be small in cognitively impaired samples, we have previously shown that utilizing that information to change MCI classification increases diagnosis accuracy and leads to earlier detection of decline (Goldberg et al., 2015; Jutten et al., 2020; Sanderson-Cimino et al., 2021).

MCI classification methods, particularly in research, almost always rely on use of cut-off scores to define cognitive impairment (Jak et al., 2009; Winblad et al., 2004). The same cut-off is typically applied at baseline and follow-up visits. If an individual with MCI at baseline experiences a PE greater than their cognitive decline, then they may be pushed above the threshold for impairment despite having no change or even a slight decline in their actual cognitive ability. Even if there were no change in cognitive capacity, this individual would likely be misclassified as CN at follow-up, appearing to revert when in fact they still have MCI. The impact of PEs on MCI reversion rates has not been explicitly studied, but it is often suggested when reversion rates are discussed (Malek-Ahmadi, 2016; Thomas et al., 2020).

Present study

In the present analyses, we utilized a sample of Alzheimer's Disease Neuroimaging Initiative (ADNI) participants who were diagnosed as MCI at baseline. We sought to 1) calculate 1-year follow-up cognitive classifications using PE-unadjusted and PE-adjusted scores, 2) compare reversion rates and diagnostic stability between PE-unadjusted and PE-adjusted classifications, and 3) provide criterion validity for the PE-adjusted classifications through baseline biomarker data and time until first dementia diagnosis. We hypothesized that the PE-adjusted scores would reveal false reverters, i.e., participants at follow-up who were classified as CN via PE-unadjusted scores but MCI via PE-adjusted scores. By retaining these participants in the MCI pool, we expected the PE-adjusted classifications to result in improved diagnostic stability and decreased reversion rates. Also, we expected the biomarker profile and the time until first dementia diagnosis of the false reverters to be more similar to the stable MCI participants than to true reverters (i.e., individuals classified as CN at follow-up based on both PE-adjusted and PE-unadjusted scores). Finally, in a post-hoc analysis, we modeled the impact of PE adjustment on studies concerned with progression to dementia, a common outcome in clinical drug trials and research studies.

MATERIALS AND METHODS

PARTICIPANTS

Data used in the preparation of this article were obtained from ADNI (adni.loni.usc.edu). The ADNI, led by Principal Investigator Michael W. Weiner, MD, was launched in 2003 as a public-private partnership. 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, see www.adni-info.org. Participants from the ADNI-1, ADNI-GO, and ADNI-2 cohorts were included.

MCI was diagnosed using the Jak-Bondi approach (Bondi et al., 2014; Edmonds et al., 2018b; Jak et al., 2009). Participants were classified as single domain MCI (amnestic, dysexecutive, or language-impaired) if their scores on 2 tests within the same cognitive domain were both greater than 1 SD below normative means. They were diagnosed as multi-domain MCI if they met the criteria for single domain MCI in more than one cognitive domain (e.g., impaired on both memory tasks and language tasks). The Jak-Bondi approach to MCI classification is favorable when compared with Petersen criteria with regard to the likelihood of progression to dementia, reversion rates, and proportion of biomarker-positive cases (Bondi et al., 2014; Edmonds et al., 2018b).

We identified 344 individuals who were classified as MCI at baseline. Of those 344, 329 returned for a 12-month follow-up visit and also completed all cognitive measures at both assessments. Mean educational level of returnees was 16.4 years (SD=2.9), 61.4% (n=202) were female, and mean baseline age was 73.1 years (SD=7.4).

PROCEDURES

Six cognitive tests were examined across the approximately 12-month test-retest interval. Episodic memory tasks included the Wechsler Memory Scaled-Revised, Logical Memory Story A delayed recall, and the Rey Auditory Verbal Learning Test (AVLT) delayed recall. Language tasks included the Boston Naming Test and Animal Fluency. Attention-executive function tasks

were Trails A and Trails B. The American National Adult Reading Test provided an estimate of premorbid IQ. Only participants who had complete test data and completed the same version of tests at the baseline and 12-month visits were included.

Z-scores were calculated for the PE-adjusted and -unadjusted scores based on independent external norms that accounted for age, sex, and education for all tests except the AVLT(Shirk et al., 2011). The AVLT was z-scored based on the ADNI participants who were CN at baseline (n=889) because we were unable to find appropriate external norms for this sample that also accounted for age, sex, and education. AVLT demographic corrections were based on a regression model that followed the same approach as the other normative adjustments. Beta values were multiplied by an individual's corresponding age, sex, and education. The products were then removed from the AVLT raw scores. These adjusted AVLT scores were then z-scored.

Baseline biomarkers included cerebrospinal fluid amyloid-beta (Aβ), phosphorylated tau (p-tau), and total tau (t-tau). The ADNI biomarker core (University of Pennsylvania) used the fully automated Elecsys immunoassay (Roche Diagnostics). Sample collection and processing have been described previously.(Shaw et al., 2009) Cutoffs for biomarker positivity were: Aβ+: Aβ<977 pg/mL; p-tau+: p-tau>21.8 pg/mL; t-tau+: t-tau>270 pg/mL (http://adni.loni.usc.edu/methods) (Elman, Panizzon, Gustavson, Franz, Sanderson-Cimino, Lyons, & Kremen, 2020; Hansson et al., 2018). There were 226 returnees with biomarker data.

Dementia was diagnosed according to ADNI criteria: 1. Memory complaint by subject or study partner that is verified by a study partner; 2. Mini-Mental State Examination score between 20-26 (inclusive); 3. Clinical Dementia Rating score of either .5 or 1; 4. An impaired delayed memory score on the Logical memory test: ≤ to 8 for 16 or more years of education; ≤ to 4 for 8-

15 years of education; or ≤ to 2 for 0-7 more years of education; 5. National Institute of Neurological and Communicative Disorders and Stroke–Alzheimer's Disease and Related Disorders Association criteria for probable AD (Petersen et al., 2010). No participants met these criteria at baseline or at the 12-month follow-up.

REPLACEMENT-PARTICIPANTS APPROACH TO PRACTICE EFFECTS

Although review papers have noted that PEs can exist even when there is longitudinal decline in observed performance, as expected within a sample at risk for AD (Salthouse, 2010), few have empirically demonstrated that claim (Goldberg et al., 2015). In such situations, Calamia et al. suggested that the most suitable approach is to utilize replacement participants (Calamia et al., 2012; Rönnlund & Nilsson, 2006). To our knowledge, the replacement-participant approach has only been utilized in two samples (Elman et al., 2018; Ronnlund et al., 2005). In this method new participants are recruited for testing at follow-up who are demographically matched to returnees. The only difference between the groups is that replacements are taking the tests for the first time whereas returnees are retaking the tests. As age is one of the matching factors, any age-related decline should be equal across the groups. Therefore, comparing scores at follow-up between returnees and replacement participants (with additional adjustment for attrition effects) allows for detection of PEs when observed scores remain stable and—unlike other methods—even when they decline. In both scenarios, scores would have been lower without repeated exposure to the tests (Elman et al., 2018; Ronnlund et al., 2005).

The goal of the replacement method is to obtain follow-up scores at retest that are free of PEs and comparable to normative data (which assume no presence of PEs). Some researchers have used PEs in other ways, such in short-term retest paradigms (Duff, 2014; Duff et al., 2014;

Duff & Hammers, 2020; Duff et al., 2011). The goal of this approach is to predict future decline and the likelihood of progressing to MCI or dementia (Jutten et al., 2020). Rather than predict decline, the goals of the replacement method are: 1) to detect decline at a given point in time that has been masked due to PEs, and 2) to revise the diagnosis of CN or MCI based on cognitive scores that have been appropriately adjusted to reflect the estimated magnitude of masked decline. Furthermore, only the replacement method has been empirically shown to calculate PEs when there is observable decline over time (Calamia et al., 2012; Elman et al., 2018). This attribute of the method makes it uniquely appropriate for samples that are impaired at baseline and/or are expected to decline over time (Calamia et al., 2012). Also, unique to this method is the fact that it allows for a change in how early MCI may be diagnosed.

PRACTICE EFFECT CALCULATION

Because replacement participants were not part of the original ADNI study design, we created what we refer to as the pseudo-replacement method of PE adjustment. We have fully described this method previously in an examination of individuals who were cognitively normal at baseline (Sanderson-Cimino et al., 2021). Briefly, a bootstrap approach (5,000 resamples, with replacement) was used to calculate PE values for each cognitive test. At every bootstrap iteration, a subsample of returnees was randomly selected (25% of sample) from the total number of individuals who had a baseline and 12-month follow-up visit. . We then removed these selected returnees from the overall baseline pool, leaving a subset of potential "pseudo-replacement participants" that included returnees not chosen at that iteration and those who did not return for a follow-up (approximately 75% of the sample). From this potential replacement pool, a set of pseudo-replacements was matched to selected returnees on age at returnee follow-up, sex, years of education, and premorbid IQ using one-to-one matching and propensity scores

(R package: MatchIt) (D. Ho, Imai, King, Stuart, & Whitworth, 2018). Additional t-tests and chi-squared tests ensured that returnees and pseudo-replacements were matched at a group level (ps> .8). Thus, this sample of pseudo-replacement participants was demographically identical to the returnee subsample. In a traditional replacement participants method of PE-adjustment returnees and non-returnees are combined into a "baseline" subsample that excludes replacements. In this method, we used a "proportional baseline" subsample that included the baseline scores for the returnees chosen at that iteration as well as all other subjects not chosen to be pseudo-replacements (approximately 75% of sample). However, the removal of the pseudo-replacements from the sample led to an artificially high portion of lower-performing baseline participants since the pseudo-replacements perform at a similar level to returnees at baseline. To correct for this issue, we calculated the retention and attrition rates for that visit in the overall sample. Because the PE for each test was calculated individually, we used test-specific retention and attrition rates, which resulted in a slight variation in rates; the average retention rate was 66% (65-70%) and the average attrition rate was 34% (30-35%). We then used these rates in the creation of the proportional baseline mean (see below). Of note, due to the bootstrapping and matching procedure, the number of participants in each group (i.e., returnees, replacements) varied but was always greater than 80 participants.

The equations below were used to calculate the PE:

$$\text{Difference score} = \text{Returnees}_{T2} - \text{Pseudo-Replacements}_{T1}$$

$$\text{Attrition effect} = \text{Returnees}_{T1} - \text{Proportional Baseline}_{T1}$$

$$\text{Practice effect} = \text{Difference score} - \text{Attrition Effect}$$

Where Returnees_{T2} represents the mean score of the returnee sample at their second

assessment, Pseudo-replacements$_{T1}$ represents the mean score of the pseudo-replacement sample (by definition, at their first assessment), and Returnees$_{T1}$ represents the mean score of returnees at their first assessment. The Proportional Baseline$_{T1}$ was a weighted mean calculated by multiplying the returnee baseline scores by the test-specific retention rate (65-75%) and the remaining portion of the subsample by the test-specific attrition rate (30-35%%). The difference score represents the sum of the PE and the attrition effect. The attrition effect accounts for the fact that individuals who return for follow-up are typically higher-performing or healthier than those who drop out. Subtracting the attrition effect from the difference score prevents over-estimation of the PE (Elman et al., 2018; Ronnlund et al., 2005). Use of a proportional baseline that retains the test-specific retention and attrition rates prevents overestimation of the attrition effect as removing the pseudo-replacements from this sample artificially lowers the baseline mean score. The PE for each test was calculated by subtracting the attrition effect from the difference score.

STATISTICAL ANALYSIS

After calculation, the PE for each test was then subtracted from each individual's observed (unadjusted) follow-up test score to provide PE-adjusted raw scores. Cohen's d was calculated for each PE by comparing PE-unadjusted and PE-adjusted scores. Adjusted raw scores at follow-up were converted to z-scores, which were used to determine PE-adjusted diagnoses. Stated differently, a score was labeled as impaired if the follow-up PE-adjusted score was greater than 1 SD below the average demographic-corrected mean. To evaluate the impact PE-adjustment had on cognitive classification, McNemar χ^2 tests were used to compare differences in the proportion of individuals classified as having MCI before and after adjusting for PEs. To assess criterion validity of the PE-adjusted diagnoses, McNemar χ^2 tests were used to compare

the number of biomarker-negative reverters and biomarker-positive stable MCI participants when using PE-adjusted versus PE-unadjusted scores.

Time until first dementia diagnosis in months from baseline was also used to validate PE-adjusted diagnoses. Cognitive data used to diagnose dementia by ADNI were not adjusted for PEs. Wilcoxon rank sum tests were used to compare groups due to the non-normal distribution of months until first dementia diagnosis. It was expected that those who reverted to CN status at follow-up would progress to dementia more slowly than those who remained classified as having MCI. As such, if PE adjustment improved diagnostic accuracy by correctly relabeling some false reverter (based on PE-unadjusted scores) as MCI, then a comparison between MCI and CN groups should show a larger and more statistically significant difference when using PE-adjusted scores than when using PE-unadjusted scores. PE-adjustment should also alter a comparison between those who truly revert and the false reverters, with false reverters progressing faster than true reverters. The following four time-until-dementia comparisons were tested: PE-adjusted MCI versus PE-adjusted CN; PE-unadjusted MCI versus PE-unadjusted CN; False reverters versus PE-unadjusted MCI; and False reverters versus PE-adjusted CN.

We also expected that the false reverters (based on PE-unadjusted scores) would have a biomarker profile more similar to the stable MCI participants than the true reverters. Thus, we calculated rates of biomarker positivity for diagnostic groups (Stable MCI and reverters) first using PE-unadjusted scores and then with PE-adjusted scores.

In post-hoc analyses, Cox proportional hazard models compared progression to dementia between those who were diagnosed as MCI at follow-up and those who reverted to CN. All models used classification (Stable MCI vs reverters) as the independent variable of interest and months from baseline until first dementia diagnosis as the dependent variable. Covariates were

age and education. Models were completed first with PE-unadjusted scores and then with PE-adjusted scores.

Time-to-dementia analyses included a full model and three timeframe-restricted models: 16-150 months (full sample data), 16-24 months, 16-36 months, and 16-48 months. The models with restricted timeframes attempted to demonstrate how predictive the classification was for studies with shorter follow-up periods. Because, in these hypothetical studies, we could not know if a participant progressed to dementia past the specified timeframe, each model was right-censored with time to event defined as time to first dementia diagnosis or time to last follow-up within the restricted time period. As this project utilized existing data, the maximum follow-up period was set to 150 months because that was the longest available timeframe within ADNI.

RESULTS

PEs were non-zero for 5 of the 6 measures (Table 1) and ranged in magnitude (Cohen's d=.06 to .26). PE-adjustment resulted in 23 more participants (+9%) classified as MCI at 1-year follow-up than when using PE-unadjusted scores (272 vs 249). Of the 23, 16 (+9%) were classified as single-domain MCI and 7 participants classified as multi-domain MCI (+9%). Regarding specific cognitive domains, PE-adjustment resulted in 24 more participants (+11%) classified with memory impairment (233 vs 209), 6 more participants (+9%) classified with attention-executive impairments (73 vs 67), and 5 more participants (+7%) classified with language impairments (72 vs 67). Full results are presented in Table 2.

The overall 1-year stability of MCI (lack of reversion to CN) was raised by 7% when adjusting for PEs (PE-adjusted stability rate=82.7%; PE-unadjusted stability rate=75.6%). Across groups (single-domain MCI, multi-domain MCI) and within each cognitive domain

(memory, attention-executive, language), PE adjustment increased the number of participants who retained their baseline diagnosis of MCI (Range: +2 [+3%] to +22 [+11%]). In particular, there were significantly more participants who remained in the impaired range at follow-up on memory when using PE-adjusted data versus PE-unadjusted data (+11%; 201 vs 223). A similar significant result was also found when considering stability of single-domain MCI (+12%; 147 vs 164). Table 3 provides full stability results.

The overall reversion rate (i.e., being classified as CN at follow-up) was 24.3% (n=80) using PE-unadjusted scores and 17.3% (n=57) using PE-adjusted scores. This indicates that adjusting for PEs resulted in a 28.8% reduction in the overall reversion rate. Table 4 describes how PE adjustment affects reversion rates across diagnostic subgroups and cognitive domains. Among those with single-domain MCI at baseline, adjusting for PEs reduced reversion rates by 27.4% (53 vs 73 reverters). Regarding specific cognitive domains, adjustment reduced the reversion rate among those with baseline memory impairments by 33.3% (44 vs 66). Adjustment also decreased reversion rates among the remaining cognitive domains (attention-executive and language) as well as among those who were multi-domain MCI at baseline (reversion to CN rate reduction range: 6.5% to 13.3%), but this equated to only a small change in the number of participants (ns<5).

We also compared how PE-adjusted and PE-unadjusted classification affected rate of progression to dementia. Of the 329 returnees, 159 progressed to dementia (48% of sample). As shown in Table 5, those who were diagnosed as MCI at follow-up and progressed to dementia during the study were first diagnosed in approximately the same time frame, regardless of PE consideration (median=25.0 months). Those who reverted to CN and later progressed to dementia did somewhat more slowly than the stable MCI groups (PE-unadjusted median=37.3

months; PE-adjusted median=60.3 months). In PE-unadjusted groups, based on Mann-Whitney U tests, there was no significant difference in time until first dementia diagnosis between stable MCI and reverter participants (W=1703; p=.177). However, in the same comparison based on PE-adjusted scores, those in the stable MCI group progressed significantly faster than those who reverted to CN (W=1240; p=.017).

Ten of the false reverters (6.2%) progressed to dementia. These participants progressed to dementia in a similar time frame as the those diagnosed with MCI via PE-unadjusted scores (median=30.03 months). The false reverters progressed to dementia more quickly than those who were classified as CN based on PE-adjusted scores at follow-up. There was not a significantly different rate of progression to dementia between false reverters and PE-adjusted CNs, or between false reverters and PE-unadjusted MCI based on Mann-Whitney U tests (*p*s>.17).

When false reverters were removed by adjusting for PEs, the median time until first dementia diagnosis was increased (+23 months). To further investigate this finding, we performed post-hoc Cox proportional hazard models to compare progression to dementia from 12-month follow-up between those who were diagnosed as MCI at follow-up and those who reverted to CN. Across all models, the hazard ratio associated with increased risk of dementia progression among stable MCI participants was nearly twice as large when adjusted for PEs compared to PE-unadjusted diagnoses (average hazard ratio: PE-adjusted=8.9, PE-unadjusted=4.2; average percent increase=110%). Figure 1 displays hazard ratios and survival curves for all models. Supplemental figure 1 provides additional Kaplan-Meier curves and risk tables for progression to dementia by diagnosis group.

There were 226 participants with baseline biomarker data. As shown in Table 6a, regardless of PE adjustment, approximately 70% of those who were diagnosed as MCI at follow-

up were Aβ positive and 70% were P-tau positive at baseline. Similarly, regardless of PE adjustment, about 60% of reverters were Aβ positive and 45% were P-tau positive. There were 18 false reverters with biomarker data. The false reverter group had an Aβ positivity of 55% and a P-tau positivity of 40%. Table 6b displays the biomarker positivity rates for each classification group based on amyloid and P-tau positivity (i.e., A-/T-, A+/T-, A-/T+, and A+/T+). Regarding the false reverters, 72% (13/18) were positive for at least one biomarker.

DISCUSSION

The validity and utility of MCI criteria are weakened by high reversion rates, which have been a longstanding problem for MCI as a construct (Pandya et al., 2016). As a result, some practitioners are hesitant to use MCI as an early indicator of AD, despite the field's goal of identifying and treating those on the AD trajectory as early as possible (Alexander et al., 2021; Canevelli et al., 2016; Pandya et al., 2016; R. A. Sperling et al., 2014). Among individuals in the ADNI sample who were diagnosed with MCI at baseline, adjusting for PEs led to a significant reduction in reversion to CN over 1 year (28.8% reduction in reversion rate). This meant that classifications were more stable across time, particularly for those with baseline amnestic MCI.

Pathologically, AD is characterized by a progressive change in amyloid beta and tau protein levels in the brain (Anand et al., 2017). Although there is conflicting evidence regarding the temporal staging of AD biomarkers and cognitive symptoms (Braak, Thal, Ghebremedhin, & Del Tredici, 2011; Emily C. Edmonds et al., 2015; Elman, Panizzon, Gustavson, Franz, Sanderson-Cimino, Lyons, Kremen, et al., 2020; C. R. Jack, Jr. et al., 2013; Veitch et al., 2019), it is likely that in most cases abnormal levels of amyloid beta are first reached, followed by abnormal levels of tau, which in turn affect cognition (Dubois et al., 2016; Jack Jr et al., 2018; Jack Jr et al., 2017). In our analyses, approximately half of the false reverters were amyloid

positive while around a third were tau positive. Nearly three-quarters of the false reverters were positive for at least one of the two biomarkers. A comparison across all three groups – true reverters, false reverters, and stable MCI – suggests that the false reverters may be an intermediate/mixed biomarker group. Some of the false reverters who were biomarker negative (A-/T-) may have MCI that is unrelated to AD. However, it is also possible that even some of the false reverters who were biomarker negative may still be on the AD trajectory. We previously showed, for example, that after controlling for tau, cognitive function in A- individuals in the ADNI sample predicted progression to A+ status (Elman, Panizzon, Gustavson, Franz, Sanderson-Cimino, Lyons, & Kremen, 2020). Overall, the PE-adjustment reduced the number of reverters, resulting in more stable MCI diagnoses and may be identifying more people who are beginning to show clinically significant levels of AD biomarkers.

Use of a robust normal sample partially addresses PEs as the cut-off for MCI diagnosis varies at each timepoint based on the distribution of scores among participants who remain CN across all visits (Emily C Edmonds et al., 2015; J. S. Eppig et al., 2017; Thomas, Edmonds, Delano-Wood, & Bondi, 2017; Thomas et al., 2019). In a similar ADNI subsample, use of robust norms found a one-year reversion rate of 15.8% (Thomas et al., 2019), which is similar to the rate found in the present study (17.3%). Whether the rates would be similar in different studies remains an open question. Using robust normals instead of normative data means that gauging impairment is based on what is a "super-normal" group that is, essentially, by definition, non-representative. This non-representativeness will be compounded further if the sample itself is not representative. For example, the robust normal group in ADNI is the highest functioning subgroup of what is already a very highly educated sample. In this approach there is no accounting for how PEs may be affecting classification into the robust normal group itself. It is

possible that some individuals in that group might actually be classified as having MCI at some follow-up if their scores were adjusted for PEs at each time point based on a replacement participants approach. Moreover, PE estimation can be overestimated if attrition effects are not considered (Elman et al., 2018; Ronnlund et al., 2005). PEs based on a robust normal group may be inflated as compared to PEs within the overall sample because, by definition, this group does not have attrition (J. S. Eppig et al., 2017; Thomas et al., 2017). Finally, comparison of results from the present study with that of our prior study (Sanderson-Cimino et al., 2021) shows that it is important to differentiate the cognitive status of individuals at baseline because the magnitude of PEs differs for individuals who are CN at baseline versus those who have MCI at baseline.

Proponents of MCI as a diagnostic entity note that individuals with the diagnosis are more likely to progress to AD, and do so at a faster rate than CN individuals (Mitchell & Shiri-Feshki, 2009; Pandya et al., 2016). Those critical of MCI's validity note that, while MCI is associated with AD, individuals with MCI are more likely to revert to CN over time than to progress to AD (Canevelli et al., 2016). Here we found that the false reverters progressed to dementia at approximately the same rate as individuals who were classified as MCI at both time points. In contrast, those who were classified as CN (i.e., true reverters) at follow-up progressed to dementia more slowly than the false reverters. These results are consistent with the notion that misclassification of these false reverters, caused by the failure to account for PEs, is weakening the predictive ability of MCI. This point is echoed by the time-to-dementia diagnosis of the reverter group. Removing the false reverters from the reverter group increased the time until first dementia diagnosis among those classified as CN by almost 2 years (37.28 versus 60.28 months).

Although adjusting for PEs slightly altered the median time until first dementia diagnosis, statistical comparisons between groups were nonsignificant. To further investigate these

findings, we completed Cox proportional hazard models. Using PE-unadjusted data, we found that the stable MCI group converted to dementia significantly faster than the (false) reverter group, as expected. When models were completed with PE-adjusted data, we found that the hazard ratios sharply increased, suggesting that the PE-adjusted classifications improved differentiation between the (true) reverters and the stable MCI participants. Not accounting for PEs may thus obscure true effects or push significance above threshold, influencing subsequent interpretation.

Interestingly, hazard ratios were less different between PE-adjusted and PE-unadjusted models when analyses were completed over the full 150-month timeframe (HRs: 6.0. versus 3.7) compared to shorter time frames (24-month HRs: 8.9 versus 3.6; and 36-month HRs: 11.6 vs 4.8). These results are consistent with the idea that PE adjustment leads to earlier detection of at-risk participants, which would be particularly important for studies with shorter follow-up periods. Importantly, clinical drug trials for AD typically involve shorter follow-up periods, so increasing the number of individuals expected to progress to dementia during the trial period will increase sensitivity to treatment effects. Therefore, failure to account for PEs may have a large impact on the design of treatment studies and interpretation of their results. Earlier detection of at-risk individuals is also of obvious importance for clinical care.

STRENGTHS AND LIMITATIONS

All participants completed the logical memory test at a screening assessment, baseline, and 12-month visit; all other tests were completed only twice. Therefore, it is possible that the PE for logical memory is misestimated. However, as the effect size of the logical memory PE is similar to that of the other memory task (AVLT), it seems likely that our estimate is still valid.

Our time until dementia analyses did not account for death. Of the 329 participants included in these analyses, 33 passed away before study completion (10.0%). The modal time until death was 48-months past baseline visit (n=8; 24% of deaths). Importantly, all participants who passed away were diagnosed as stable MCI (impaired at baseline and follow-up) by both the PE-adjusted and PE-unadjusted datasets. As such, although mortality may have impacted results, this effect was equal within the PE-adjusted and PE-unadjusted analyses.

The ADNI sample was not designed to be a population-representative study. It represents a population of older adults likely to volunteer for clinical trials, and consists primarily of white, highly educated individuals who may be at a higher genetic risk for dementia than typical Americans. Results of the present study may not be applicable to other studies with different sample characteristics or retest intervals. Additionally, age and education have been shown to impact PEs (Calamia et al., 2012; Gross et al., 2017). We strongly believe that the exact PE values found in this study should not be applied to other samples, particularly if they involve CN individuals with different demographics (i.e., age and education). However, a strength of the replacement-participants method of estimating PEs is that it is always tailored to the sample, including age and education, as well as the retest interval being studied. For example, in addition to the 1-year interval in the present study, the replacement-participants method has been used successfully in studies with intervals as long as 5-6 years (Elman et al., 2018; Ronnlund et al., 2005). Participant demographics and cognitive tests are always matched. Retest intervals may vary across studies, but PEs are calculated for the specific interval(s) used within a given study. Therefore, we explicitly recommend against using these PE estimates in other studies. Rather we encourage others to utilize the method within their study to more accurately generate PEs given their specific demographics, measures, and test-retest interval. The cost of including replacement

participants might seem prohibitive, but it is actually a relatively small component in a large-scale study (Elman et al., 2018; Sanderson-Cimino et al., 2021). Elsewhere, we have shown that it could save millions of dollars in a large clinical trial because MCI is detected earlier, resulting in reductions in study duration and necessary sample size (Sanderson-Cimino et al., 2021) . As shown in the present study, the method can be adapted to large studies that did not include replacements in their original design. However, building it into the original study design is clearly preferable.

CONCLUSIONS

Here we have shown that a replacement method of PE adjustment significantly altered how we understand follow-up status in individuals who have already been diagnosed with MCI at the baseline assessment. Our results indicate that the replacement-participants method of adjustment for PEs results in fewer MCI cases reverting to CN, and improved predictability of progression to dementia. In sum, the results provide further support for the importance of accounting for PEs on cognitive tests in order to reduce misdiagnosis and increase earlier detection of progression to MCI or dementia.

Acknowledgments: The content of this article is the responsibility of the authors and does not necessarily represent official views of the National institute of Aging or the Department of Veterans affairs. The ADNI and funding sources had no role in data analysis, interpretation, or writing of this project. The corresponding author was granted access to the data by ADNI and conducted the analyses. The study was supported by grants from the U.S. National institute on Aging (MSC: F31AG064834, WSK, CEF, MJL: R01 AG050595, CEF, WSK: P01 AG055367; WSK, R01 AG022381, AG054002, AG060470; CEF, R01 AG059329; ALG: K01 AG050699; MWB: R01 AG049810; KRT: R03AG070435) and the National Center for Advancing Translational Sciences (JAE: KL2 TR001444). The Center for Stress and Mental Health in the Veterans Affairs San Diego Healthcare System also provided support for this study.

Author contributors: The study was conceived by MSC and WSK. Guidance on statistical analysis was provided by XMT and ALG. Determination of MCI diagnoses was made by ECE, MWB, JE, KRT. MSC, WSK, JAE, MSP, and DEG contributed to the practice effects methodology. Primary funding to support this work was obtain by WSK, CEF, MJL, and MSC. All authors provided critical review and commentary on the manuscript. Data collection and sharing for this project was funded by the Alzheimer's Disease

Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company

Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

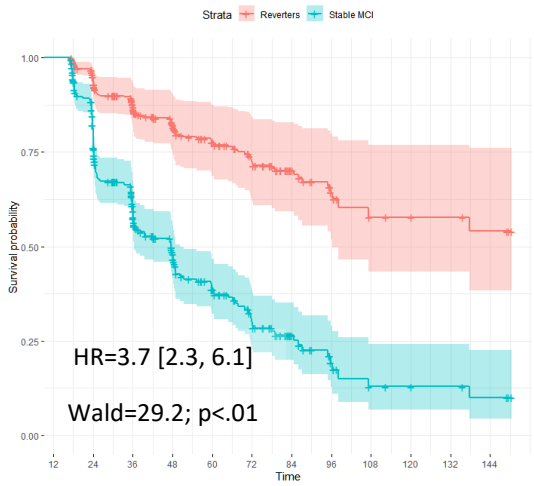
Contribution to the Field: Studies of aging and Alzheimer's disease (AD) rely on serial cognitive testing to determine change in cognitive abilities. AD clinical trials also rely on repeated testing to determine drug effects. However, scores on these tests can increase over time due to practice effects (PEs), which can increase scores at follow-up compared to if the participant was taking the test for the first time. PEs are especially impactful on studies that differentiate between unimpaired participants and those with mild cognitive impairment (MCI) because PEs can lead to misdiagnoses. Here we demonstrated that adjusting for PEs resulted in more stable MCI diagnoses, and may be identifying people who are beginning to show clinically significant levels of AD biomarkers. We also demonstrated that adjusting for PEs improves the relationship between MCI and conversion to dementia. Identifying those at risk for cognitive decline as early as possible may maximize opportunities for intervention to slow disease

progression. The results of this study demonstrate the importance of accounting for PEs on cognitive tests in order to reduce misdiagnosis and increase earlier detection of progression to MCI or dementia.

Potential conflict of interests: Dr. Bondi receives royalties from Oxford University Press. All other authors declare no competing interests.

Tables and Figures

PE-unadjusted



PE-adjusted

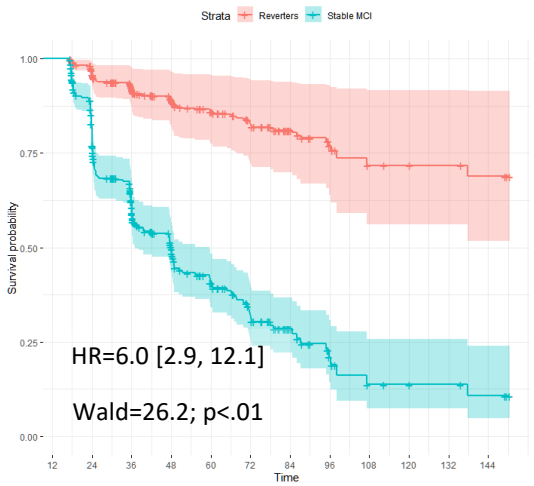


Figure 6: Full Cox proportional models for time until first dementia diagnosis by PE-unadjusted and PE-adjusted 12-month diagnoses. Cox proportional hazard models compared progression to dementia between those who were classified with mild cognitive impairment at follow-up (Stable MCI) and those who reverted to cognitively normal (Reverters). Models used classifications (Stable MCI vs Reverter) as the independent variable of interest; months from baseline until first dementia diagnosis as the dependent variable; and all available data (16 – months form baseline). Covariates were age and education, fixed at the average level within the sample (age: 73.1 years; education: 16.4 years). The left graph bases diagnoses on the PE-unadjusted 12-month data; the right graph uses diagnoses based on the PE-adjusted 12-month data. Each model presents a hazard ration (HR; [CI]) that indicates how much more likely the Stable MCI group was to convert to dementia compared to the Reverters. Wald tests and likelihood-ratio tests (LRT) are also included with associated p-values to denote the significance of the HR. The Y-axis of each model provides the survival probability and the X-axis of each model provides the time frame until dementia conversion.

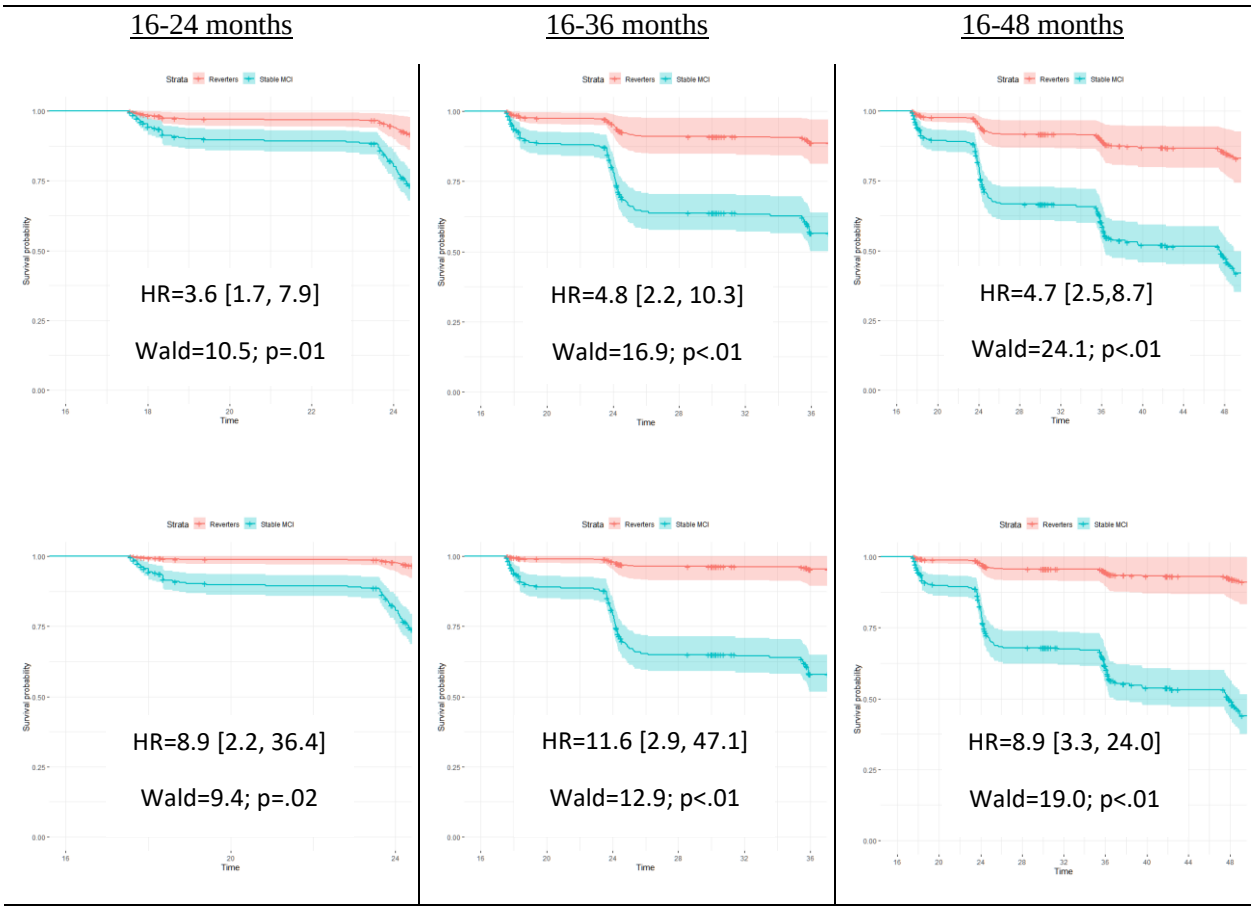


Figure 7: Cox proportional models for time until first dementia diagnosis by PE-unadjusted and PE-adjusted 12-month diagnoses. Cox proportional hazard models compared progression to dementia between those who were classified with mild cognitive impairment at follow-up (Stable MCI) and those who reverted to cognitively normal (Reverters). All models used classifications (Stable MCI vs Reverter) as the independent variable of interest and months from baseline until first dementia diagnosis as the dependent variable. Covariates were age and education, fixed at the average level within the sample (age: 73.1 years; education: 16.4 years). Models in the top row display results completed with PE-unadjusted scores; models in the bottom row display results completed with the PE-adjusted scores. Each row designates the time frame for each model measured in months from baseline. Time frames were restricted to demonstrate how predictive the classification was for studies with various follow-up periods. As these hypothetical studies would not know if a participant converted to dementia past their follow-up period, those who converted after the endpoint of that specific model were censored (i.e., recoded as non-converters). Each model presents a hazard ration (HR; [CI]) that indicates how much more likely the Stable MCI group was to convert to dementia compared to the Reverters. Wald tests and likelihood-ratio tests (LRT) are also included with associated p-values to denote the significance of the HR. The Y-axis of each of the 6 models provides the survival probability and the X-axis of each model provides the time frame until dementia conversion.

Table 3a: Descriptive statistics among participants at baseline and 1-year-follow-up

	Memory		Attention/Executive Function		Language	
Raw mean score (SD)	**RAVLT**	**Logical Memory**	**Trails A**	**Trails B**	**Boston Naming**	**Category Fluency**
Full Sample Baseline	1.55 (2.61)	5.81 (3.57)	39.27 (20.85)	106.14 (66.90)	27.82 (3.76)	15.88 (4.76)
Full Sample Follow-up	2.17 (3.09)	6.39 (4.55)	39.39 (20.67)	106.44 (74.67)	28.15 (4.10)	15.29 (5.51)

The "Full Sample" rows refer to the means (standard deviations) of all participants at baseline and at follow-up. Presents values in raw units.

Table 3b: Descriptive statistics and calculated practice effects for tests among participants classified as mild cognitive impairment at baseline

	Memory		Attention/Executive Function		Language	
Raw mean score (SD)	**RAVLT**	**Logical Memory**	**Trails A**	**Trails B**	**Boston Naming**	**Category Fluency**
Proportional Baseline	1.59 (2.61)	1.92 (3.68)	40.28 (22.75)	109.76 (75.03)	27.66 (4.16)	15.51 (4.82)
Returnees Baseline	1.58 (2.61)	2.00 (3.56)	39.88 (21.73)	107.45 (68.16)	27.77 (3.94)	15.70 (4.81)
Returnees Follow-Up	2.45 (3.07)	2.84 (4.51)	39.30 (22.19)	107.73 (76.53)	28.11 (4.51)	15.02 (5.46)
Replacements Follow-Up	1.67 (2.57)	1.86 (3.72)	41.35 (22.63)	114.40 (74.90)	27.37 (4.51)	15.11 (4.81)
Attrition Effect	-.01 [-.13, .16]	.09 [-.10, .43]	-.40 [-1.57, .89]	-2.31 [-6.64, 2.27]	.11 [-.14,.33]	.43 [.15, .72]
Practice Effect	.80 [-.33, 3.08]	.89 [-.41, 3.33]	-1.64 [-5.65, 2.41]	-4.36 [-19.16, 9.57]	.63 [-.21,1.53]	NA
Cohen's d	.26	.20	-.07	-.06	.14	NA

Groups are based on the average performance across all 5000 bootstrapped iterations. Means are based on transformed data that was reverted back to raw units. "Proportional baseline" refers to a weighted mean that combines the returnee baseline group and a group that included all subjects not selected to be Returnees or Replacements in that bootstrapped iteration. "Returnee Baseline" refers to baseline test scores for the subset of participants who returned for the 12-month follow-up visit (ns>80) and were selected at that iteration. "Returnee Follow-Up" refers to 12-month scores for the same subset of returnees who were selected for that iteration. "Replacement Follow-up" refers to the pseudo-replacement scores (ns>80). The scores for memory tasks indicate the number of words remembered at the delayed recall trials. Scores on the attention/executive functioning tests indicate time to completion of task. On these tasks, higher scores indicate worse performance. Scores on the Boston Naming Task indicate number of correct items identified; scores on Category Fluency indicate number of items correctly stated. Practice effects and attrition effects are in raw units with the 2.5 and 97.5 percentiles in brackets. As such, the negative practice effects and attrition effects for the Trails tasks demonstrates that practice decreased time (increased performance). Cohen's d is given for the difference between PE-adjusted and unadjusted scores of returnees at follow-up. RAVLT= Rey Auditory Verbal Learning Test.

Table 4: Classification prevalence at baseline and follow-up.

	Any MCI	M MCI	S MCI	Memory Impairment	Attention/EF Impairment	Language Impairment	CN
Baseline	329	75	254	267	77	70	0
Unadjusted	249	79	170	209	67	67	80
Adjusted	272	86	186	233	73	72	57
Difference	+23	+7	+16	+24	+6	+5	-23
% Difference	9.23%	8.86%	9.41%	11.48%	9.00%	7.46%	28.75%
χ^2; ***p-value***	21.0; p<.001	5.1; p=.02	7.5; p=.006	22.0; P<.001	3.2; p=.07	3.2; p=.07	21.0; p<.001

Presents the number of participants who met criteria for mild cognitive impairment (MCI). The "unadjusted" and "adjusted" rows refer to diagnoses at the follow-up visit. The "Any MCI" column presents the count of participants who meet criteria for MCI in any domain, combining those who are impaired in only one domain (single-domain MCI: S MCI) and those who are impaired in 2 or 3 domains (multiple-domain MCI: M MCI). The impairment columns present the count of participants who were impaired in each domain, regardless of whether they are impaired in another domain. Individuals who do not meet criteria for impairment (i.e., classified as Cognitively Normal; CN) are displayed in the "CN" column

The Difference row displays how many more participants meet criteria for that classification or impairment when adjusting for practice effects (i.e., Adjusted count – Unadjusted count). The percent listed in this row displays the percent increase/decrease when accounting for practice effects: difference/Unadjusted count. McNemar χ^2 tests were used to evaluate the impact of practice-effect adjustment on classification or impairment count; p-values are presented.

Table 5: Impact of practice effects on classification stability and progression

	Stable M MCI	Stable S MCI	Progression to M MCI	Stable Impairment		
				Memory	Attention/EF	Language
Unadjusted	45	147	34	201	46	42
Adjusted	49	164	37	223	48	44
Difference	+4	+17	+3	+22	+2	+2
% Difference	8.89%	11.56%	8.82%	10.94%	4.35%	4.76%
χ^2;	2.25;	11.13;	1.3;	20.0;	.5;	.5;
p-value	p=.13	p<.001	p=.25	p<.001	p=.48	p=.48

Displays the number of individuals classified as impaired at follow-up via practice effect-unadjusted scores and -adjusted scores. The "Stable M MCI" column provides the count of participants who met criteria for multiple domain mild cognitive impairment (M MCI) at baseline and at follow-up. The "Stable S MCI" provides the same information about individuals with single domain MCI (S MCI). Individuals who progressed from S MCI at baseline to M MCI at follow-up are displayed in the "Progression" column. The "Stable Impairment" section describes the number of individuals who retained an impairment in a specific cognitive domain at follow-up, regardless of whether they met criteria for an impairment in another domain at either visit. The Difference row displays how many more participants meet criteria for that classification or impairment when adjusting for practice effects (i.e., Adjusted count – Unadjusted count). The percent listed in this row displays the percent increase in stability when accounting for practice effects: difference/Unadjusted count. McNemar χ^2 tests were used to evaluate the impact of practice-effect adjustment on classification or impairment stability; p-values are presented.

Table 6: Practice effect-adjustment and reversion rates

	Reverters M MCI	Reverters S MCI	Reversion in specific domain		
			Memory	**Attention/EF**	**Language**
Count					
Unadjusted	30	73	66	28	31
Adjusted	26	53	44	26	29
Difference	-4	-20	-22	-2	-2
χ^2; ***p-value***	2.25	18.1	20.0	.5	.5
	p=.13	p<.001	p<.001	p=.48	p=.48
Reversion Rate					
Unadjusted	40.5%	28.7%	24.7%	36.3%	44.3%
Adjusted	35.1%	20.9%	16.5%	33.8%	41.4%
Difference	-5.4%	-7.8%	-8.2%	2.6%	2.9%
% change in reversion	Δ13.3%	Δ27.4%	Δ33.3%	Δ7.1%	Δ6.5%

The "**Count**" section displays the number of participants who reverted from a classification or impairment based on practice effect-unadjusted and -adjusted data. Those who reverted from multi-domain mild cognitive impairment (M MCI) at baseline to either single domain MCI (S MCI) or cognitively normal are displayed in the "Reverters M MCI" column. Those who were classified as S MCI at baseline and reverted to cognitively normal at follow-up are listed in the "Reverters S MCI" column. The "Reversion in Specific Domain" section refers to individuals who had a baseline impairment in a domain (memory, attention/executive functioning, or language) but not at follow-up; participants in these columns may be impaired in other domains at either baseline or follow-up. The Difference row displays how many fewer participants reverted when adjusting for practice effects (i.e., Adjusted count – Unadjusted count). McNemar χ^2 tests were used to evaluate the impact of practice-effect adjustment on classification or impairment reversion; p-values are presented.

The "**Reversion Rate**" section lists the reversion percent for each column by dividing the counts provided above by the baseline prevalence of each classification shown in table 1. For example, 74 people were classified as M MCI at baseline and 30 reverted at follow-up when using unadjusted data. Therefore, the reversion rate for the unadjusted M MCI reverters was 30/74. The difference row subtracts the reversion rate using Unadjusted data from the rate using Adjusted data. The "% change in reversion" row shows the percent change in reversion rate by dividing the difference by the unadjusted reversion rate: e.g., Δ13.3 =5.4/40.5.

Table 7: Progression to dementia.

Months until DX	Full Sample N=159	Stable MCI Unadjusted N=141	Stable MCI Adjusted N=151	Reverters Unadjusted N=18	Reverters Adjusted N=8	False reverters N=10
Mean	37.48	36.17	36.32	47.77	59.44	38.44
Median	25.28	24.98	24.98	37.28	60.28	30.03
SD	21.90	20.66	20.66	28.68	33.34	21.70

Presents the time in months until first dementia diagnosis (DX) among those who converted to dementia. Of the 329 participants 159 have progressed to dementia ("Full Sample"). Participants were classified as "Stable MCI" if they retained their mild cognitive impairment (MCI) classification at follow-up; participants were classified as "Reverters" if they were classified as cognitively normal at follow-up. Classifications were made using practice effect-unadjusted ("Unadjusted) and practice effect-adjusted ("Adjusted") data. Those who were classified as MCI by the practice effect-adjusted data but not the unadjusted data are referred to as "False reverters." Values are bolded to emphasize that the False reverters appear to be similar to the Stable MCI group in time to first dementia diagnosis.

Table 8a: Amyloid, total tau, and phosphorylated tau across classification groups

	Full Sample N=226	Stable MCI Unadjusted N=166	Stable MCI Adjusted N=184	Reverters Unadjusted N=60	Reverters Adjusted N=42	False Reverters N=18
Amyloid						
Count	160	124	134	36	26	10
%	70.8%	74.7%	72.8%	60.0%	61.9%	55.6%
T-tau						
Count	123	101	106	22	17	5
%	54.4%	60.8%	57.6%	36.7%	40.5%	27.8%
P-tau						
Count	145	118	125	27	20	7
%	64.2%	71.1%	67.9%	45.0%	47.6%	39.9%

Presents the number of participants (Count) and percent of sample (%) for three cerebrospinal fluid biomarkers: amyloid beta (Abeta), Tau, and phosphorylated tau (Ptau). Of the 329 participants, 226 had full biomarker data, which is presented in the "Full Sample" column. Participants were classified as "Stable MCI" if they retained their mild cognitive impairment (MCI) classification at follow-up; participants were classified as "Reverters" if they were classified as cognitively normal at follow-up. Classifications were made using practice effect-unadjusted ("Unadjusted) and practice effect-adjusted ("Adjusted") data. Those who were classified as MCI by the practice effect-adjusted data but not the unadjusted data are referred to as "False reverters." The percent sample (%) was determined by dividing the number of biomarker-positive subjects in a cell by the total number of participants with that classification; e.g., 74% = 117/158.

Table 8b: Combined Amyloid and Tau positivity profiles

	Full Sample	Stable MCI		Reverters		False Reverters n=18
		Unadjusted	Adjusted	Unadjusted	Adjusted	
A-T-						
Count	39	22	27	17	12	5
Percent	17.3%	13.3%	14.7%	28.3%	28.6%	27.8%
A+T-						
Count	42	26	32	16	10	6
Percent	18.5%	15.7%	17.4%	26.7%	23.8%	33.3%
A-T+						
Count	27	20	23	7	4	3
Percent	11.9%	12.0%	12.5%	11.7%	9.5%	16.7%
A+T+						
Count	118	98	102	20	16	4
Percent	52.2%	59.0%	55.4%	33.3%	38.1%	22.2%
A+ and/or T+						
Count	187	144	157	43	30	13
Percent	82.7%	86.7%	85.3%	71.7%	71.4%	72.2%

Presents the number of participants (Count) and percent of sample (%) for combinations of cerebrospinal fluid biomarker positivity: biomarker-negative (A-/T-), amyloid-positive and tau-negative (A+/T-), amyloid-negative and tau-positive (A-/T+), amyloid and tau positive (A+/T+), and positive for any biomarker (A+ and/or T+).

References

1 Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . 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 and Dementia, 7*, 270-279. doi:10.1016/j.jalz.2011.03.008

2 Alexander, G. C., Emerson, S., & Kesselheim, A. S. (2021). Evaluation of aducanumab for Alzheimer disease: scientific evidence and regulatory review involving efficacy, safety, and futility. *JAMA, 325*(17), 1717-1718.

3 Anand, A., Patience, A. A., Sharma, N., & Khurana, N. (2017). The present and future of pharmacotherapy of Alzheimer's disease: A comprehensive review. *European journal of pharmacology, 815*, 364-375.

4 Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., . . . Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *J Alzheimers Dis*. doi:10.3233/JAD-140276

5 Braak, H., Thal, D. R., Ghebremedhin, E., & Del Tredici, K. (2011). Stages of the Pathologic Process in Alzheimer Disease: Age Categories From 1 to 100 Years. *Journal of Neuropathology & Experimental Neurology, 70*(11), 960-969. doi:10.1097/NEN.0b013e318232a379

6 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.

7 Canevelli, M., Grande, G., Lacorte, E., Quarchioni, E., Cesari, M., Mariani, C., . . . Vanacore, N. (2016). Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis. *Journal of the American Medical Directors Association, 17*(10), 943-948.

8 Cummings, J. L., Morstorf, T., & Zhong, K. (2014). Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's research & therapy, 6*(4), 1-7.

9 Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., . . . Blennow, K. (2016). Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia, 12*(3), 292-323.

10 Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical neuropsychologist, 28*(5), 714-725.

11 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.

12 Duff, K., & Hammers, D. B. (2020). Practice effects in mild cognitive impairment: A validation of Calamia et al.(2012). *The Clinical neuropsychologist*, 1-13.

13 Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., . . . McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry, 19*(11), 932-939.

14 Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2018a). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: A secondary analysis of the ADCS vitamin E and donepezil in MCI study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*(1), 11-18.

15 Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2018b). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: A secondary analysis of the ADCS vitamin E and donepezil in MCI study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*, 11-18.

16 Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., . . . Salmon, D. P. (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia, 11*(4), 415-424.

17 Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Bondi, M. W., & Alzheimer's Disease Neuroimaging, I. (2015). Subtle Cognitive Decline and Biomarker Staging in Preclinical Alzheimer's Disease. *Journal of Alzheimer's disease : JAD, 47*(1), 231-242. doi:10.3233/JAD-150128

18 Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., . . . Jacobson, K. C. (2018). Underdiagnosis of mild cognitive impairment: A consequence of

ignoring practice effects. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*.

19 Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., & Kremen, W. S. (2020). Amyloid-β Positivity Predicts Cognitive Decline but Cognition Predicts Progression to Amyloid-β Positivity. *Biological Psychiatry*.

20 Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., . . . Initiative, A. s. D. N. (2020). Amyloid-β positivity predicts cognitive decline but cognition predicts progression to amyloid-β positivity. *Biological Psychiatry, 87*(9), 819-828.

21 Eppig, J., Werhane, M., Edmonds, E. C., Wood, L.-D., Bangen, K. J., Jak, A., . . . Bondi, M. W. (2020). Neuropsychological Contributions to the Diagnosis of Mild Cognitive Impairment Associated With Alzheimer's Disease. *Vascular Disease, Alzheimer's Disease, and Mild Cognitive Impairment: Advancing an Integrated Approach*, 52.

22 Eppig, J. S., Edmonds, E. C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M. W., & Initiative, A. s. D. N. (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.

23 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.

24 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: The Journal of the Alzheimer's Association, 13*(7), P473-P474.

25 Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., . . . 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.

26 Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J. Q., Bittner, T., . . . Alzheimer's Disease Neuroimaging, I. (2018). CSF biomarkers of Alzheimer's disease concord with amyloid-beta PET and predict clinical progression: A study of fully

automated immunoassays in BioFINDER and ADNI cohorts. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. doi:10.1016/j.jalz.2018.01.010

27 Ho, D., Imai, K., King, G., Stuart, E., & Whitworth, A. (2018). Package 'MatchIt'. In: Version.

28 Jack, C. R., Jr., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., . . . Trojanowski, J. Q. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol, 12*(2), 207-216. doi:10.1016/S1474-4422(12)70291-0

29 Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Karlawish, J. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562.

30 Jack Jr, C. R., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., Knopman, D. S., . . . Kantarci, K. (2017). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimer's & Dementia, 13*(3), 205-216.

31 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 : official journal of the American Association for Geriatric Psychiatry, 17*(5), 368-375. doi:10.1097/jgp.0b013e31819431d5

32 Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A., Jones, R. N., Choi, S. E., . . . Tommet, D. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1), e12055.

33 Malek-Ahmadi, M. (2016). Reversion from mild cognitive impairment to normal cognition. *Alzheimer Disease & Associated Disorders, 30*(4), 324-330.

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

35 Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica, 119*(4), 252-265.

36 Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., . . . Strobel, G. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology, 15*(7), 673-684.

37 Pandya, S. Y., Clem, M. A., Silva, L. M., & Woon, F. L. (2016). Does mild cognitive impairment always lead to dementia? A review. *Journal of the neurological sciences, 369*, 57-62.

38 Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., . . . Toga, A. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology, 74*(3), 201-209.

39 Rönnlund, M., & Nilsson, L.-G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence, 34*(1), 63-78.

40 Ronnlund, M., Nyberg, L., Backman, 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. *Psychol Aging, 20*(1), 3-18. doi:10.1037/0882-7974.20.1.3

41 Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society, 16*(5), 754-760.

42 Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., . . . Eppig, J. S. (2021). Cognitive Practice Effects Delay Diagnosis; Implications for Clinical Trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, Preprint*.

43 Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., . . . Alzheimer's Disease Neuroimaging, I. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol, 65*(4), 403-413. doi:10.1002/ana.21610

44 Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy, 3*(6), 32.

45 Sperling, R., Mormino, E., & Johnson, K. (2014). The evolution of preclinical Alzheimer's disease: implications for prevention trials. *Neuron, 84*(3), 608-622.

46 Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: stopping AD before symptoms begin? *Science translational medicine, 6*(228), 228fs213-228fs213.

47 Thomas, K. R., Cook, S. E., Bondi, M. W., Unverzagt, F. W., Gross, A. L., Willis, S. L., & Marsiske, M. (2020). Application of neuropsychological criteria to classify mild cognitive impairment in the active study. *Neuropsychology, 34*(8), 862.

48 Thomas, K. R., Edmonds, E. C., Delano-Wood, L., & Bondi, M. W. (2017). Longitudinal trajectories of informant-reported daily functioning in empirically defined subtypes of mild cognitive impairment. *Journal of the International Neuropsychological Society, 23*(6), 521-527.

49 Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., . . . Salmon, D. P. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia, 15*(10), 1322-1332.

50 Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., . . . Morris, J. C. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia, 15*(1), 106-152.

51 Wang, G., Kennedy, R. E., Goldberg, T. E., Fowler, M. E., Cutter, G. R., & Schneider, L. S. (2020). Using practice effects for targeted trials or sub-group analysis in Alzheimer's disease: How practice effects predict change over time. *PloS one, 15*(2), e0228064. doi:10.1371/journal.pone.0228064

52 Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., . . . Almkvist, O. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of internal medicine, 256*(3), 240-246.

Chapter 2, in full, is a reprint of the material as it appears in Fronters of Aging Neuroscience. Mark Sanderson-Cimino, Jeremy A. Elman, Xin M. Tu, Alden L. Gross, Matthew S. Panizzon, Daniel E. Gustavson, Mark W. Bondi, Emily C. Edmonds, Joel S. Eppig, Carol E. Franz, Amy J. Jak, Michael J. Lyons, Kelsey R. Thomas, McKenna E. Williams, and William S. Kremen, PhD. *Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments.* Frontiers in Aging Neuroscience, 2022. **14**. The dissertation author was the primary investigator and author of this paper.

Chapter 3: Misinterpreting change over multiple timepoints: When cognitive practice effects meet age-related decline

Title: Misinterpreting cognitive change over multiple timepoints: When practice effects meet age-related decline

Authors: Mark Sanderson-Cimino[1,2*], Ruohui Chen[3]*, Xin M. Tu[4,5,6], Jeremy A. Elman[2,8], Amy J. Jak[1,2,7], William S. Kremen[1,2,4]

Affiliations: [1]Joint Doctoral Program in Clinical Psychology, San Diego State/University of California, San Diego; [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

*Joint first authors

Corresponding author: Mark Sanderson-Cimino, Department of Psychiatry and Center for Behavior Genetics of Aging, University of California, San Diego, 9500 Gilman Dr. (MC 0738), La Jolla, CA 92093.

E-mail: mesander@health.ucsd.edu

Objective: Practice effects (PE) on cognitive testing 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, the participant-replacement method can separate pathology- or age-related decline from PEs; however, this method has only been used across two timepoints. More than two timepoints makes 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 1190 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 the 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.

Conclusions: The 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.

Public Significance: PEs mask decline on cognitive tests. If they are not considered at each visit, then clinicians may delay treatment or diagnosis of impairment; researchers may inaccurately label longitudinal trends. Unlike other methods, the replacement method demonstrates that PEs may be present even when performance declines. In contrast to a commonly held view, PEs also do not necessarily level off after the first follow-up. Specifically, PEs impacted multiple follow-up visits and failure to consider them led to inaccurate conclusions about change.

Keywords: Practice effects, cognitive aging, longitudinal change.

Introduction

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; J. Eppig et al., 2020; 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; J. Eppig et al., 2020; 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 the first time (Calamia et al., 2012; Heilbronner et al., 2010; Salthouse, 2019; Sanderson-Cimino 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., 2017; 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 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 post-baseline (Elman et al., 2018; Goldberg et al., 2015; Gross et al., 2017; Gross et al., 2015; Rönnlund, Nyberg, Bäckman, & Nilsson, 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., 2017; Gross et al., 2015; Sanderson-Cimino et al., 2021; Sanderson-Cimino et al., 2022). It has been suggested that addressing PEs might lead to earlier diagnosis of impairment (Goldberg et al., 2015; Sanderson-Cimino et al., 2021), 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 earlier detection of MCI, improves stability of MCI diagnoses, and strengthens our ability to predict conversion to dementia (Elman et al., 2018; Sanderson-Cimino et al., 2021; Sanderson-Cimino 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 (Elman et al., 2018; Sanderson-Cimino et al., 2021; Sanderson-Cimino 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 (e.g., older adults and, particularly, those on the Alzheimer's disease [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., 2014; Duff et al., 2011). In contrast, 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 point here is that there is still a PE despite the fact that performance declines, because the 1-point decline is masking a true 2-point decline. 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 et al., 2021). Moreover, studies with multiple time points typically conclude that the magnitude of PEs 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, leading to less of an observable increase in scores. To our knowledge, this hypothesis has been raised, but never formally tested.

In the present analyses we used a modified version of the -replacement method of PE adjustment (see Methods) (Elman et al., 2018; Rönnlund et al., 2005; Sanderson-Cimino et al., 2021; Sanderson-Cimino et al., 2022) and generalized estimating equations (GEE) to examined 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 6 neuropsychological measures. Models were completed first without adjusting for PEs and second after adjusting for PEs. We hypothesized that: (1) PEs will be present at both the 12-month follow-up and the 24-month follow-up; (2) PEs will increase across time; (3) the PE-adjusted models will find both significant age-related decline and significant PEs; (4) PE-unadjusted models will provide inaccurate estimates of change as compared to the PE-adjusted models.

Methods

Participants: All participants were enrolled in the ADNI (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 (MRI), positron emission tomography (PET), 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, see www.adni-info.org. Participants from the ADNI-1, ADNI-GO, and ADNI-2 cohorts were included.

There were 1190 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 one SD below normative means. They were diagnosed as multi-domain 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, Bornstein, & Prifitera, 1990), and the Rey Auditory Verbal Learning Task (AVLT) delayed recall (Schmidt, 1996). Language tasks included the Boston Naming Test (BNT) (Kaplan, Goodglass, & Weintraub, 2001) and Category (Animal) Fluency (Petersen et al., 2010). Attention-executive function tasks were Trails A and Trails B (Lezak, Howieson, Loring, & Fischer, 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: We developed a pseudo-replacement method of PE adjustment to adapt the replacement method for use in studies that did not specifically

recruit replacement participants. This method relies on the identification of a subsample of baseline participants who are demographically matched to the returnees at follow-up and labeled as "pseudo-replacements." Propensity scores are used to ensure that the pseudo-replacements and returnees are similar, with respect to age and other demographic characteristics. A comparison of the pseudo-replacement 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 replacements have not. As such, pseudo-replacements are functionally the same as replacement participants who are recruited for that purpose as part of a study's original design. The replacement method has been used for PEs at a single retest visit (Elman et al., 2018; Rönnlund et al., 2005; Sanderson-Cimino et al., 2021; Sanderson-Cimino 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 pseudo-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=257). The remaining 70% include dropouts and returnees who did not attend all visits. This group served as a pool for potential pseudo-replacements to calculate PEs at the 12-month follow-up. Propensity scores were used to select replacements that matched these returnees on age, sex, and ANART scores. We define this group of replacements as Group A; at their baseline they had similar demographics to the subsampled returnees at their 12-month follow-up. To estimate the PEs at 24-month follow up, we followed a similar strategy, and used the same

subsampled 30% of participants' follow up data at the 2-year follow-up timepoint. Pseudo-replacements were matched to these returnees on age, sex, and the ANART using propensity scores. We define this group of replacements as Group B; they had completed tests at the 12-month follow and have similar demographics to the subsampled 30% participants at their 24-month follow-up. 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.

After creating the matched pseudo-replacement groups' data, we used GEE to estimate the time effects and PEs. Let 't' denote the timepoint at which participants take the test; 't=1' represents the baseline, 't=2' represents 12-month follow up and 't=3' represents 24-month follow-up. 'G' denotes group, and let 'G=R' represent 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:

$$E[Y_{it}|\boldsymbol{X}_{it}] = \beta_0 + \boldsymbol{\beta}_x\boldsymbol{X}_{it} + \beta_1 I(t=2) + \beta_2 I(t=3)$$

$$+ \beta_3 I(t=2)I(G=A) + \beta_4 I(t=3)I(G=B)$$

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 second visit and 691 (85%) returned for a third visit. Within the group that was MCI at baseline (n=381), 381 (100%) returned for a second visit and 241 (63%) returned for a third 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. PE estimates did not uniformly change over time. PEs on the Trails tests reduced over time (Trails A: -11%; Trails B: -26%), as did PEs on the AVLT (-5%). However, PEs increased over time on Logical Memory (+39%), BNT (+41%), and category fluency (+9%; nonsignificant 24-month PE). The PEs in raw score units are fully presented in Table 2.

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. 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 patten of change over time was significantly different in the PE-adjusted and PE-unadjusted groups. Figure 1 displays the trajectories for PE-adjusted and PE-unadjusted scores for all cognitive measure among participants who were CU at baseline.

Logical Memory: 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 follow-up. In contrast, when PE-estimates were included in the model, scores significantly worsened across time (-.54 and -2.0, respectively). The PE-unadjusted and the PE-adjusted age effects had non-overlapping 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: The PE-unadjusted model indicated nonsignificant improved performance over time (+.71, +.32). In contrast, the PE-adjusted model indicated significantly worsening performance over time (-.30, -.61). The PE-unadjusted and the PE-adjusted age effects had non-overlapping confidence intervals at the 24-month follow-up ([-.73, 1.4] vs [-.90, -3.2])

BNT: The PE-unadjusted model indicated nonsignificant improvement in scores over time (+44, +.08). The PE-adjusted model showed nonsignificant improvement at the 12-month follow up (+.02) but significantly worse scores at the 24-month follow-up (-.56).

Category fluency: The PE-unadjusted model indicated nonsignificant improvement in scores at the 12-month follow-up (+.72) and significant improvement in scores at the 24-month follow-up (+.40). In the PE-adjusted model there was nonsignificant worsening in scores at the 12-month follow-up (-.19) and significant worsening of scores at the 24-month visit (-.60).

Trails A: The PE-unadjusted model indicated nonsignificant improvement in scores at the 12-month follow-up (+2.72) and significant improvement in scores at the 24-month follow-up (+1.29). In contrast, when adjusting for PEs, there was nonsignificant worsening of scores at the 12-month visit (-.99) and significant worsening of scores at the 24-month follow-up (-2.23). The PE-unadjusted and the PE-adjusted age effects had non-overlapping 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: The PE-unadjusted model indicated nonsignificant improvement in scores at the 12-month follow-up (+11.6) and at the 24-month follow-up (+6.87). The PE-adjusted model indicated nonsignificant worsening in scores at the 12-month follow-up (-1.42) and a significant worsening of scores at the 24-month follow-up (-4.0). The PE-unadjusted and the PE-adjusted age effects had non-overlapping 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 2 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 (+.79) and at the 24-month follow-up (+2.4). The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up (-.72) and at the 24-month follow-up (-1.0). The PE-unadjusted and the PE-adjusted age effects had non-overlapping 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 (+.07) and at the 24-month follow-up (-.09). The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up (-.52) and 24-month follow-up (-.89). The PE-unadjusted and the PE-adjusted age effects had non-overlapping 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 (+1.3) and nonsignificant worsening of scores at the 24-month follow-up (-.58). The PE-adjusted model showed nonsignificant improvement in scores at the 12-month follow-up (+.24) and at the 24-month follow-up (+.36).

Category Fluency: In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up (+.99) and nonsignificant worsening of scores at the 24-month follow-up (-2.0). The PE-adjusted model showed significant worsening of scores at the 12-month follow-up (-.72) and at the 24-month follow-up (-1.01).

Trails A: In the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up (+5.96) and nonsignificant worsening of scores at the 24-month follow-up (-.92). The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up (-.57) and a near-zero change at the 24-month follow-up (.00).

Trails B: Within the PE-unadjusted model, there was nonsignificant improvement in scores at the 12-month follow-up (+26.30) and at the 24-month follow-up (+30.01). The PE-adjusted model showed nonsignificant worsening of scores at the 12-month follow-up (-3.24) and a nonsignificant improvement in scores at the 24-month follow-up (+1.24).

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 magnitude 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, de Jager, Hogervorst, & Budge, 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 groups or it could potentially 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 affected 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 significant PEs that are likely impacting incidence rates of MCI and stability of MCI diagnoses within ADNI (Sanderson-Cimino et al., 2021; Sanderson-Cimino 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 et al., 2021), whether one (Petersen et al., 2010) or more than one impaired measure (Jak et al., 2009) is required.

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 two 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 non-significant in the PE-adjusted model. However, the confidence intervals were non-overlapping 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. Therefore, adjusting for PEs led to results that more accurately represent the expected cognitive change.

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 to improve the accuracy of cognitive data and diagnoses (e.g., earlier diagnoses, more stable diagnoses) (Elman et al., 2018; Sanderson-Cimino et al., 2021; Sanderson-Cimino 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 of much aging research, these small PE-adjustments can have significant downstream effects on everything from the utility of biomarkers to study duration (Sanderson-Cimino et al., 2021; Sanderson-Cimino et al., 2022).

PEs have been shown to weaken over time in some studies (Machulda et al., 2017; Stricker et al., 2020), leading many to conclude that consideration of PEs is less important at subsequent follow-up visits (Goldberg et al., 2015; Heilbronner et al., 2010; 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 et al., 2021). Salthouse et al 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 the 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 study of older adults with multiple follow-ups over a more than four-year period found that PEs largely leveled out after the first follow-up assessment (Stricker et al., 2020). As is also common, 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 well-done study. However, as with many studies, we think their 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 present with age-related decline, PEs may exist even when the actual scores decrease. The advantage of the replacement method is that accounting for age-related declines 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 is not restricted to very old adults. For example, Elman et al. (Elman et al., 2018) demonstrated the same phenomenon in a six-year follow-up of middle-aged adults in their mid-50s at baseline.

In the clinic, neuropsychologists do not have the potential benefit of replacement participants as may be the case in research. 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), 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 six or 12 months).

Strengths and Limitations

Participants for this project were drawn from ADNI, which consisted primarily of highly educated, healthy, white individuals who typically presented to memory clinics (Petersen et al., 2010). The magnitude of PEs may be different in other samples or individuals with different backgrounds or demographics. If a sample differs in the characteristics of its participants -for example has an average age of 40 years versus this sample's 73 years– PE estimates may differ considerably. Therefore, we note that estimates of PEs from this study should not be applied to other samples. Instead, we recommend that researchers utilize this method to create PEs specific to their sample. While the exact PEs should not be used by other studies, a major strength of the replacement method is that it always produces tailor-made PE estimates because returnees and replacements are always matched on their unique features including demographics, specific tests, and retest intervals.

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 et al., 2021; Sanderson-Cimino 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.

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 or who have been diagnosed with amnestic MCI (Eppig et al., 2017; Sanderson-Cimino et al., 2021; Sanderson-Cimino et al., 2022; Thomas et al., 2019). Future studies with larger sample sizes may benefit from conducting sub-analyses 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.

Tables and figures

Table 9: Sample demographics and raw cognitive scores

	Baseline Age	Baseline ANART	Education
CN	73.70 (6.79)	10.16 (7.94)	16.38 (2.64)
MCI	73.07 (7.34)	13.42 (9.64)	15.98 (2.83)

Baseline

	n	LM	AVLT	BNT	CF	TA	TB
CN	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)

12-month follow-up

	n	LM	AVLT	BNT	CF	TA	TB
CN	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)

24-month follow-up

	n	LM	AVLT	BNT	CF	TA	TB
CN	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)

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. The American National Adult Reading Test (ANART) was given at baseline only and the provided scores are the total errors on the test. 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 leftmost column. **LM**-Logical Memory; **AVLT**-Rey Auditory Verbal Learning Task; **BNT**-Boston Naming Test; **CF**-Category Fluency; **TA**-Trails A; **TB**-Trails B.

Table 10: Estimates for generalized estimating equation models within the subsample diagnosed as cognitively unimpaired at baseline

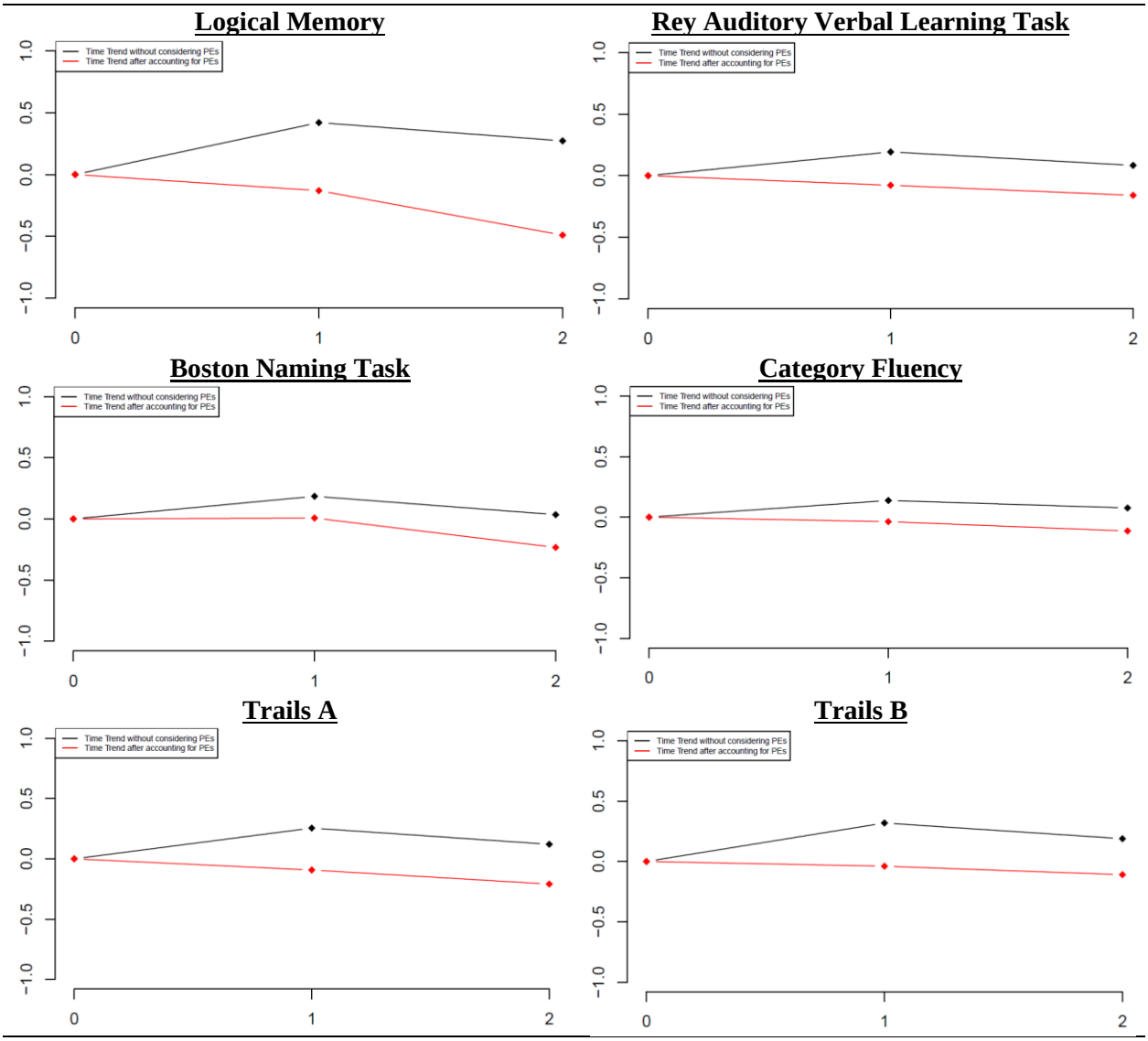
	12 month follow-up			24 month follow-up		
	Unadjusted Age-Effect	**Practice Effect**	**Adjusted Age-Effect**	**Unadjusted Age-Effect**	**Practice Effect**	**Adjusted Age-Effect**
LM	+1.74* [1.4,2.1]	+2.28* [1.5,3.1]	-.54* [-.84,-.24]	+1.13* [.83,1.4]	+3.18* [2.2,4.2]	-2.0* [-2.3,-1.7]
AVLT	+.74 [-.11,1.6]	+.98* [.38,1.6]	-.30* [-.57,-.02]	+.32 [-.73,1.4]	+.93* [.02,1.8]	-.61* [-.90,-3.2]
BNT	+.44 [-.12,1.0]	+.46* [.09,.84]	+.02 [-.13,.16]	+.08 [-.46,.63]	+.65* [.14,1.2]	-.56* [-.73,-.39]
CF	+.72 [-.44,1.9]	+.96* [.07,1.9]	-.19 [-.54,.16]	+.40* [.05,.75]	+1.05 [-.25,2.4]	-.60* [-.97,-.23]
Trails A	+2.72 [-.11,5.5]	+3.81* [1.8,5.8]	-.99* [-1.8, -.22]	+1.29* [.32,2.3]	+3.40* [.74, 6.1]	-2.23* [-3.1,-1.4.]
Trails B	+11.6 [-1.8,25.0]	+12.94* [5.5,20.4]	-1.42 [-4.4,1.6]	+6.87 [-.07,13.8]	+9.59* [-.15,19.3]	-4.0* [-7.1,-.82]

Presents beta 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 as 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 "*".Gray highlighting designates non-overlapping 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.

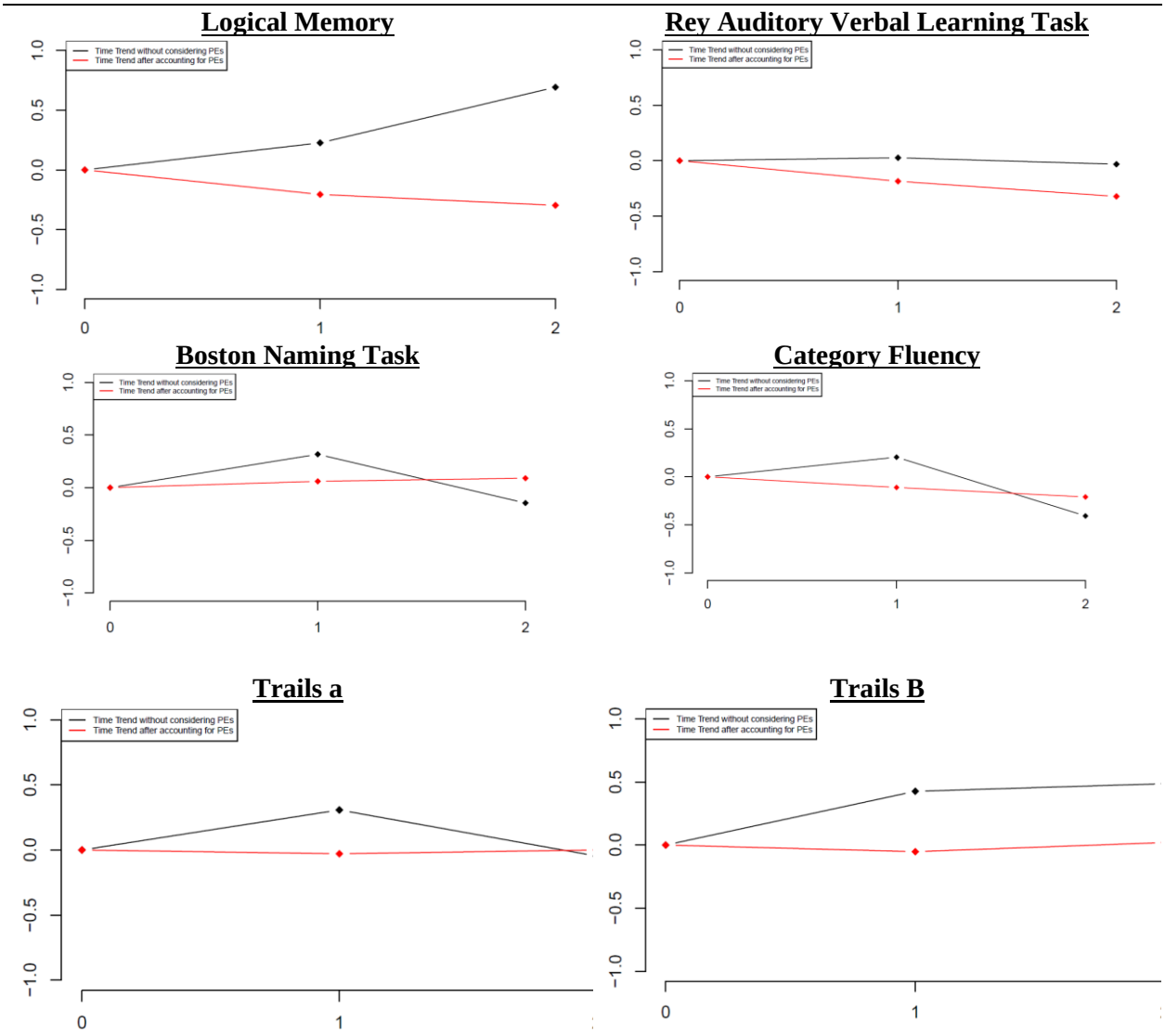
Table 11: Practice effect and time estimates for generalized estimating equations within a subsample diagnosed with mild cognitive impairment at baseline

	12 month follow-up			24 month follow-up		
	Unadjusted Age-Effect	**Practice Effect**	**Adjusted Age-Effect**	**Unadjusted Age-Effect**	**Practice Effect**	**Adjusted Age-Effect**
LM	+.79 [-1.9,3.5]	+1.50* [.84,2.2]	-.72 [-1.1,-.35]	+2.4 [-.11,4.9]	+3.50* [1.9,5.1]	-1.0* [-1.5,-.52]
AVLT	+.07 [-.21,.35]	+.62* [.15,1.1]	-.52 [-.80,-.24]	-.09 [-.44,.26]	+.80 [-.16,1.8]	-.89 [-1.3,-.53]
BNT	+1.3 [-.30,2.8]	+1.01 [.26,1.78]	+.24 [-.14,.62]	-.58 [-2.5,1.3]	-.92 [-2.7,.86]	+.36 [-.09,.81]
CF	+.99* [-9.6,2.9]	+1.50 [.62,2.4]	-.54* [-1.0,-.03]	-2.0 [-4.4,.49]	-.80 [-3.0,1.4]	-1.01* [-1.6,-.38]
Trails A	+5.96 [-1.6, 13.5]	+6.68* [2.2,11.1]	-.57 [-2.8,1.7]	-.92 [-4.0,2.2]	-.90 [-15.6,13.8]	.00 [-2.9,2.9]
Trails B	+26.30 [-5.2, 57.8]	+35.34* [21.7,49.0]	-3.24 [-10.0,3.6]	+30.01 [-1.2,61.2]	+30.66* [2.2,59.1]	+1.24 [-6.9,9.3]

Presents beta 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 as 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 "*".Gray highlighting designates non-overlapping 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.



Graph 1: Expected cognitive scores among participants who were unimpaired at baseline. The Y-axis of each graph presents standardized scores for all 6 cognitive measures. The X-axis indicates the baseline (0), 12-month follow-up (1) and 24-month follow-up (2). The dark line provides estimates from the practice effect-unadjusted generalized estimating equation model; the red line presents estimates from the practice effect-adjusted model. 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.



Graph 2: Expected cognitive scores among participants diagnosed with mild cognitive impairment at baseline. The Y-axis of each graph presents standardized scores for all 6 cognitive measures. The X-axis indicates the baseline (0), 12-month follow-up (1) and 24-month follow-up (2). The dark line provides estimates from the practice effect-unadjusted generalized estimating equation model; the red line presents estimates from the practice effect-adjusted model. 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.

References

1 Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . 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 and Dementia, 7*, 270-279. doi:10.1016/j.jalz.2011.03.008

2 Byrd, D. A., & Rivera-Mindt, M. G. (2022). Neuropsychology's race problem does not begin or end with demographically adjusted norms. *Nature Reviews Neurology*, 1-2.

3 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.

4 Chelune, G. J., Bornstein, R. A., & Prifitera, A. (1990). The Wechsler memory scale—revised. In *Advances in psychological assessment* (pp. 65-99): Springer.

5 Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical neuropsychologist, 28*(5), 714-725.

6 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.

7 Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., . . . McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry, 19*(11), 932-939.

8 Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., . . . Jacobson, K. C. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*.

9 Eppig, J., Werhane, M., Edmonds, E. C., Wood, L.-D., Bangen, K. J., Jak, A., . . . Bondi, M. W. (2020). Neuropsychological Contributions to the Diagnosis of Mild Cognitive Impairment Associated With Alzheimer's Disease. *Vascular Disease, Alzheimer's Disease, and Mild Cognitive Impairment: Advancing an Integrated Approach*, 52.

10 Eppig, J. S., Edmonds, E. C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M. W., & Initiative, A. s. D. N. (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.

11 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.

12 Galvin, J. E., Powlishta, K. K., Wilkins, K., McKeel, D. W., Xiong, C., Grant, E., . . . Morris, J. C. (2005). Predictors of preclinical Alzheimer disease and dementia: a clinicopathologic study. *Archives of neurology, 62*(5), 758-765.

13 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.

14 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: The Journal of the Alzheimer's Association, 13*(7), P473-P474.

15 Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., . . . 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.

16 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.

17 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. *American Journal of Geriatric Psychiatry, 17*, 368-375. doi:10.1097/JGP.0b013e31819431d5

18 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 : official journal of the American Association for Geriatric Psychiatry, 17*(5), 368-375. doi:10.1097/jgp.0b013e31819431d5

19 Jutten, R. J., Thompson, L., Sikkes, S. A., Maruff, P., Molinuevo, J. L., Zetterberg, H., . . . Gold, M. (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*(Preprint), 1-14.

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

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

22 Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., . . . 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.

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

24 Mathews, M., Abner, E., Kryscio, R., Jicha, G., Cooper, G., Smith, C., . . . 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.

25 Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., . . . Toga, A. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology, 74*(3), 201-209.

26 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.

27 Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International neuropsychological Society, 16*(5), 754-760.

28 Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and aging, 34*(1), 17.

29 Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., . . . Eppig, J. S. (2021). Cognitive Practice Effects Delay Diagnosis;

Implications for Clinical Trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, Preprint.*

30 Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. 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*. doi:10.3389/fnagi.2022.847315

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

32 Schrijnemaekers, A., de Jager, C. A., Hogervorst, E., & Budge, 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.

33 Stricker, N. H., Lundt, E. S., Alden, E. C., Albertson, S. M., Machulda, M. M., Kremers, W. K., . . . 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.

34 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.

35 Thomas, K. R., Cook, S. E., Bondi, M. W., Unverzagt, F. W., Gross, A. L., Willis, S. L., & Marsiske, M. (2020). Application of neuropsychological criteria to classify mild cognitive impairment in the active study. *Neuropsychology, 34*(8), 862.

36 Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., . . . Salmon, D. P. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia, 15*(10), 1322-1332.

37 Vivot, A., Power, M. C., Glymour, M. M., Mayeda, E. R., Benitez, A., Spiro III, A., . . . 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.

Chapter 3, in part, is under peer review. The dissertation author was the primary investigator and author of this paper.

7. Integrated discussion

PEs interfere with our ability to detect changes in cognitive functioning. The overall goal of this dissertation was to investigate how PE-adjustment through the participant replacement method improves accuracy of longitudinal cognitive assessment. *Paper 1* applied an adapted form of the replacement method for use in studies that did not specifically recruit matched replacement participants. The primary aims of this paper were to: use the method to calculate PEs in older adults who were CU at baseline; compare MCI incidence rates based on PE-adjusted and PE-unadjusted scores; and to validate the PE-adjusted MCI diagnoses. A secondary aim of this paper was to demonstrate how PE-adjustment could lead to increased efficiency as well as monetary savings in a hypothetical AD drug trial. *Paper 2* expanded upon the results in paper 1 by applying the method in individuals with MCI at baseline. The goals of the project were to: quantify PEs in a sample of impaired individuals; assess how PEs impact MCI reversion rates and stability; and to determine if PE-adjustment alters the association between diagnostic group and later dementia. *Paper 3* combined GEE models and the pseudo-replacement approach to simultaneously calculate decline over time and PEs across 3 visits (baseline, 1 year, 2 years). The goals of this project were to: determine if PEs are significant across multiple time points; assess if PEs decline across visits when simultaneously modeled with change over time; and to determine if PE adjustment significantly alters estimates of decline over time.

The first project found non-zero PEs using the pseudo-replacement method across a 1-year retest interval on 5 of 6 cognitive measures. Although these PEs were small in magnitude (Cohen's d range: .08 to .24), accounting for PEs when making follow-up diagnoses led to a significant increase in the incidence of MCI (+19%; 104 vs 124). Unlike prior studies, I was then able to validate these results through biomarker analyses. I demonstrated that there was an

increased proportion of amyloid-positive MCI cases (+14%; 51 vs 58) when using PE-adjusted vs PE-unadjusted follow-up scores. Moreover, after adjusting for PEs, there was a decrease in amyloid-positive CU subjects (-5%; 152 vs 145). These biomarker analyses provided necessary criterion validity for diagnoses created with PE-adjusted data.

Using recruitment data from the Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4) Study I demonstrated that adjusting for PEs can result in substantial financial savings -in the range of several millions of dollars- for an AD clinical drug trial even with the additional cost required for replacement participants. More than just a budgetary issue, these results showed that studies and clinical trials that incorporate the replacement method of PE adjustment can increased their statistical power, partly because they result in earlier detection of MCI. This substantially reduced the number of individuals that are needed for initial screening, amyloid PET, and enrollment in the study. Earlier detection of MCI incidence also means that study duration can be reduced, which would meaningfully reduce staff and subject burden, and might also reduce dropout rates. In summary, the first project contributed to the literature by validating PE-adjusted MCI diagnoses via biomarker positivity and showing that the replacement method leads to earlier detection of progression to MCI. It also provided quantitative estimates of the beneficial effects that this would have on studies with progression to MCI as an outcome. In addition, it also demonstrated the viability of a pseudo-replacement method of PE-adjustment.

The second paper expanded the analyses to a subsample of ADNI participants that were already diagnosed with MCI at baseline. Similar to the first paper, I found non-zero PEs for 5 of 6 measures (Cohen's d range: .06-.26). Adjusting for PEs led to a significant increase in the proportion of participants diagnosed as MCI at follow-up as compared to diagnoses based on PE-unadjusted scores (+9%; 249 vs 272). As all participants were diagnosed with MCI at baseline,

retention of MCI at follow-up meant a reduction in reversion rates (-29%; 57 vs 80 reverters). This implies that adjusting for PEs improved the stability of the MCI diagnoses, addressing a primary concern of those critical of MCI as a diagnostic entity (Canevelli et al., 2016; Pandya et al., 2016). Skeptics of MCI also comment on how reversion rates impact progression to dementia analyses (Canevelli et al., 2016; Pandya et al., 2016). I found that false reverters (those diagnosed as CU by PE-unadjusted scores but MCI by PE-adjusted scores) progressed to dementia at approximately the same rate as individuals who were classified as MCI at both time points (i.e., stable MCI). In contrast, those who were classified as CU based on PE-adjusted diagnoses at follow-up (i.e., true reverters) progressed to dementia more slowly than the false reverters. These results are consistent with the notion that misclassification of these false reverters, caused by the failure to account for PEs, is weakening the predictive ability of MCI. Similar to the first paper, I also conducted post-hoc analyses to highlight the importance of adjusting for PEs. I compared hazard models to investigate how PE adjustment improves the association of MCI with later progression to dementia. Our models indicate that use of PE-adjusted scores at the 12-month follow-up visit resulted in an approximate 2-fold increase in hazard ratios. The greatest change in hazard ratios was seen in models covering a shorter time frame (i.e., 12-24 months from baseline) as compared to the full model (i.e., 12-150 months from baseline). This finding is significant because most clinical trials have study intervals much less than 150 months (e.g., the A4 study). In summary this project found PEs among individuals who were diagnosed with MCI at baseline. Accounting for these PEs at the 12-month follow-up visit significantly reduced reversion rates, increased MCI stability, and doubled the association with subsequent progression to dementia.

The third project expanded the pseudo-replacement method of PE-adjustment for use across multiple time points. Using GEE I found significant PEs at a 12-month and 24-month follow up visits in two subsamples (CU or MCI at baseline). Specifically, within a sample that was CU at baseline, there were significant PEs effects for all measures at the first retest (12-month visit) and significant PEs on 5 of 6 measures at the second retest (24-month visit). Within the sample that was MCI at baseline, PEs were significant for 4 of 6 measures at the first retest and 2 of 6 measures at the second retest. Importantly, using this method I found that with re-testing at 1-year intervals the PEs at the second follow-up were in some cases larger than the PEs at the first follow-up. For example, the PE on Logical Memory was about 1 point higher at the second follow-up within the CU sample (2.3 vs 3.2), and around 2 points higher in the MCI sample (1.5 vs 3.5). These results are in contrast to a well-documented argument that PEs decline over time, particularly after the first retesting (Goldberg et al., 2015; Heilbronner et al., 2010; Vivot et al., 2016). However, there are a small number of studies suggesting that this trend is not consistent (Calamia et al., 2012; Wilson, Watson, Baddeley, Hazel, & Evans, 2000) . The findings in the present study indicate that simultaneously modeling PEs and age- or pathology-related decline does help to clarify the cognitive trajectories, including trajectories of PEs.

The combination of the replacement method and GEE models allowed for simultaneous consideration of PEs and age/pathological change over time. As part of the third paper, I compared sets of analyses that included and excluded PEs to investigate how consideration of PEs impacts longitudinal studies. Among the CU baseline sample, the PE-unadjusted analyses demonstrated stable or significantly *improved* scores across both follow-ups. In contrast, the PE-adjusted models found stable or significantly *decreased* scores across visits, with a greater decline at the second retest. Therefore, I found that scores tended to stay the same or increase

when PEs were ignored. If this trend accurately reflects cognitive ability, that would mean that this sample of older adults is improving their cognitive ability over 2 years. While possible, this interpretation is highly unlikely given that the parent study, ADNI, recruited participants that were similar those in AD clinical drug trials and had a higher-than-average risk for neurodegeneration (Petersen et al., 2010). It would also mean that similar levels of performance would be observed on other tests in the same cognitive domains. However, as Goldberg et al. note, such transfer of training is not generally observed (Goldberg et al., 2015). In contrast, when including PEs, there was a negative change in scores across visits that in many cases had non-overlapping confidence intervals with analyses that did not adjust for PEs. There was considerably more heterogeneity in estimates within the MCI baseline sample, partly due to floor effects. As a result, many of the change-over-time variables were nonsignificant in the GEE models. However, the overall pattern was similar to that of the CU sample; when adjusting for PEs, there was generally a decline in scores across time while PE-unadjusted models typically found stable or improved scores across time.

Taken together, these findings highlight the importance of considering PEs, and using the replacement method, when completing longitudinal studies, particularity with populations that are expected to decline over time (Salthouse, 2010, 2019). The primary finding of this research is that this particular method of accounting for PEs leads to earlier diagnosis, even if there is a group-level decline in cognitive ability. I also found that the method applies to samples that are CU or MCI at baseline, and that accounting for PEs can improve the stability and predictive utility of MCI diagnoses. Although some prior work has investigated this method (Ronnlund et al., 2005; Sanderson-Cimino et al., 2022), this work was the first to validate the early PE-adjusted diagnoses, which was accomplished via biomarkers and progression to dementia data.

Early and valid diagnoses are paramount to our advancement of research and clinical treatment of atypical aging (Albert et al., 2011; J. Eppig et al., 2020; Manly et al., 2008; Mitchell & Shiri-Feshki, 2009; Pandya et al., 2016; Thomas et al., 2020). Importantly, to my knowledge, no other method of estimating PEs can alter how early a diagnosis can be made. With respect to research, ignoring PEs hampers our ability to separate cases from controls or to accurately track cognitive change. This type of PE-adjustment becomes a data correction tool to improve the accuracy of cognitive diagnoses and data. As diagnostic groups and cognitive scores are basic components or outcome measures of most aging research, these small PE-adjustments can have significant downstream effects on everything from biomarkers to study duration (Sanderson-Cimino et al., 2021; Sanderson-Cimino et al., 2022).

I believe that this work strongly supports the idea that it is ideal for aging studies to include replacement participants in their original design. However, I also demonstrated that the method can be adapted to large studies that did not specifically recruit replacement participants or that had multiple follow-up visits. Although this project utilized a method that is not viable for clinical care of individuals, it does suggest that clinicians should strongly consider PEs when making diagnoses. As the field considers the current state of normative data (Byrd & Rivera-Mindt, 2022) and there is a strong push for repeated assessments, particularly via computerized this dissertation testing (Jutten et al., 2022; Stricker et al., 2020), this study strongly suggests that PEs need to be considered. Moreover, I found that PEs may not decrease across time as is typically claimed in both research and clinical practice (Goldberg et al., 2015; Heilbronner et al., 2010; Vivot et al., 2016). This claim is likely based on the idea that PEs only exist if there is an *increase* in scores over time. As I have demonstrated through this series of studies, this assumption is less applicable in studies where samples are expected to decline over time. Indeed,

I provided evidence that PEs for episodic memory may actually increase after the first follow-up assessment. Therefore, it is likely to be beneficial for clinical practice to develop norms for PEs form a diverse range of individuals retested at clinically-relevant intervals (e.g., at 6- and 12-month follow-ups). Although this may be seen as a time-consuming endeavor, the findings in these 3 projects suggest that the scientific and financial benefits likely outweigh the costs.

8. Strengths and limitations

This study was completed within ADNI, a sample that is primarily of white and highly educated. They are also simultaneously healthier (e.g., cardiovascular health) and at greater risk for dementia that the typical American. Thus, the results of these studies may not translate to other studies. Additionally, PEs are known to vary based on many factors including age, education, and retest interval(Calamia et al., 2012). They also differ across assessment measures, even for those within the same cognitive domain(Elman et al., 2018). It is not recommended that the PE estimates provided by this research be used by other studies. However, this is not a limitation of the method per se because the PE adjustment methods that I have applied are designed to be utilized by any study with sufficient sample size. It is possible to match replacements on nearly any factor, including biomarkers and health factors. As such, while the exact PEs should not be used by other studies, the method is still viable because it always produces tailor-made PE estimates based on that sample's unique features. That is, the method requires that returnees and replacements are always matched on demographics, specific tests, and retest intervals. I am, for example, implementing the method in ongoing research in other samples with very different designs and demographics. I applied the method to an epidemiological, population-based cohort study focused on characterizing cardiovascular risk factors among individuals who self-identified as Hispanic or Latino in the United States

(HCHS/SOL). Participants in that study were matched at baseline on cardiovascular health and PEs were found after a 7-year retest interval. The method has also been utilized within a separate older adult sample that completed the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) annually for 5 years.

9. Conclusion

In summary, these papers comment on the importance of cognitive PEs, particularly within studies of cognitive aging. This is the first series of studies to demonstrate that the replacement method of PE-adjustment can lead to earlier diagnoses of MCI and can improve the stability of MCI diagnoses. These results validated the replacement method of PE adjustment by showing greater biomarker positivity and increased progression to dementia analyses in MCI cases. Integration of the GEE into this method revealed that PEs can actually increase after the first retesting. Also, the models using GEE confirmed that inclusion of PEs can uncover sharper age- or pathology-related decline, even if the sample appears to have worsening cognitive performances over time. These biomarker analyses, in combination with a hypothetical clinical drug trial, illustrated how PEs have significant downstream effects within any study utilizing cognitive data. The findings showed that implementing the replacement method could save millions of dollars and substantially reduce study duration as well as staff and participant burden in a clinical trial. Moving forward, this research should be replicated in a more diverse sample and include additional follow-up visits. Regarding clinical contributions, this work suggests that the field would benefit from the development of PE-adjusted normative data for clinically useful retest intervals.

References

1 2018 Alzheimer's disease facts and figures. (2018). *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 14*(3), 367-429. doi:10.1016/j.jalz.2018.02.001

2 Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., . . . 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 and Dementia, 7*, 270-279. doi:10.1016/j.jalz.2011.03.008

3 Alexander, G. C., Emerson, S., & Kesselheim, A. S. (2021). Evaluation of aducanumab for Alzheimer disease: scientific evidence and regulatory review involving efficacy, safety, and futility. *JAMA, 325*(17), 1717-1718.

4 Anand, A., Patience, A. A., Sharma, N., & Khurana, N. (2017). The present and future of pharmacotherapy of Alzheimer's disease: A comprehensive review. *European Journal of Pharmacology, 815*, 364-375.

5 Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., . . . Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *J Alzheimers Dis*. doi:10.3233/JAD-140276

6 Braak, H., Thal, D. R., Ghebremedhin, E., & Del Tredici, K. (2011). Stages of the Pathologic Process in Alzheimer Disease: Age Categories From 1 to 100 Years. *Journal of Neuropathology & Experimental Neurology, 70*(11), 960-969. doi:10.1097/NEN.0b013e318232a379

7 Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia, 3*(3), 186-191.

8 Byrd, D. A., & Rivera-Mindt, M. G. (2022). Neuropsychology's race problem does not begin or end with demographically adjusted norms. *Nature Reviews Neurology*, 1-2.

9 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.

10 Canevelli, M., Grande, G., Lacorte, E., Quarchioni, E., Cesari, M., Mariani, C., . . . Vanacore, N. (2016). Spontaneous reversion of mild cognitive impairment to normal

cognition: a systematic review of literature and meta-analysis. *Journal of the American Medical Directors Association, 17*(10), 943-948.

11 Chelune, G. J., Bornstein, R. A., & Prifitera, A. (1990). The Wechsler memory scale—revised. In *Advances in psychological assessment* (pp. 65-99): Springer.

12 Cummings, J., Lee, G., Ritter, A., Sabbagh, M., & Zhong, K. (2019). Alzheimer's disease drug development pipeline: 2019. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 5*, 272-293.

13 Cummings, J. L., Morstorf, T., & Zhong, K. (2014). Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's research & therapy, 6*(4), 1-7.

14 Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., . . . Blennow, K. (2016). Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia, 12*(3), 292-323.

15 Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical neuropsychologist, 28*(5), 714-725.

16 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.

17 Duff, K., & Hammers, D. B. (2020). Practice effects in mild cognitive impairment: A validation of Calamia et al.(2012). *The Clinical neuropsychologist*, 1-13.

18 Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., . . . McCaffrey, R. J. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *The American Journal of Geriatric Psychiatry, 19*(11), 932-939.

19 Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2018a). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: A secondary analysis of the ADCS vitamin E and donepezil in MCI study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*(1), 11-18.

20 Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2018b). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: A secondary analysis of the ADCS vitamin E

and donepezil in MCI study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, 4*, 11-18.

21 Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., . . . Salmon, D. P. (2015). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia, 11*(4), 415-424.

22 Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Bondi, M. W., & Alzheimer's Disease Neuroimaging, I. (2015). Subtle Cognitive Decline and Biomarker Staging in Preclinical Alzheimer's Disease. *Journal of Alzheimer's disease : JAD, 47*(1), 231-242. doi:10.3233/JAD-150128

23 Edmonds, E. C., Delano-Wood, L., Jak, A. J., Galasko, D. R., Salmon, D. P., & Bondi, M. W. (2016). "Missed" mild cognitive impairment: High false-negative error rate based on conventional diagnostic criteria. *Journal of Alzheimer's Disease, 52*(2), 685-691.

24 Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., . . . Jacobson, K. C. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*.

25 Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., & Kremen, W. S. (2020). Amyloid-β Positivity Predicts Cognitive Decline but Cognition Predicts Progression to Amyloid-β Positivity. *Biological Psychiatry*.

26 Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., . . . Initiative, A. s. D. N. (2020). Amyloid-β positivity predicts cognitive decline but cognition predicts progression to amyloid-β positivity. *Biological Psychiatry, 87*(9), 819-828.

27 Elman, J. A., Vuoksimaa, E., Franz, C. E., & Kremen, W. S. (2020). Degree of cognitive impairment does not signify early versus late mild cognitive impairment: confirmation based on Alzheimer's disease polygenic risk. *Neurobiology of aging, 94*, 149-153.

28 Eppig, J., Werhane, M., Edmonds, E. C., Wood, L.-D., Bangen, K. J., Jak, A., . . . Bondi, M. W. (2020). Neuropsychological Contributions to the Diagnosis of Mild Cognitive Impairment Associated With Alzheimer's Disease. *Vascular Disease, Alzheimer's Disease, and Mild Cognitive Impairment: Advancing an Integrated Approach*, 52.

29 Eppig, J. S., Edmonds, E. C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M. W., & Initiative, A. s. D. N. (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.

30 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.

31 Galvin, J. E., Powlishta, K. K., Wilkins, K., McKeel, D. W., Xiong, C., Grant, E., . . . Morris, J. C. (2005). Predictors of preclinical Alzheimer disease and dementia: a clinicopathologic study. *Archives of neurology, 62*(5), 758-765.

32 Gauthier, S., Albert, M., Fox, N., Goedert, M., Kivipelto, M., Mestre-Ferrandiz, J., & Middleton, L. T. (2016). Why has therapy development for dementia failed in the last two decades? *Alzheimer's & Dementia, 12*(1), 60-64.

33 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.

34 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: The Journal of the Alzheimer's Association, 13*(7), P473-P474.

35 Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., . . . 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.

36 Gross, A. L., Inouye, S. K., Rebok, G. W., Brandt, J., Crane, P. K., Parisi, J. M., . . . Jones, R. N. (2012). Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time. *Journal of Clinical and Experimental Neuropsychology, 34*(7), 758-772.

37 Gustavson, D. E., Elman, J. A., Sanderson-Cimino, M., Franz, C. E., Panizzon, M. S., Jak, A. J., . . . Kremen, W. S. (2020). Extensive memory testing improves prediction of progression to MCI in late middle age. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1).

38 Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J. Q., Bittner, T., . . . Alzheimer's Disease Neuroimaging, I. (2018). CSF biomarkers of Alzheimer's disease concord with amyloid-beta PET and predict clinical progression: A study of fully automated immunoassays in BioFINDER and ADNI cohorts. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. doi:10.1016/j.jalz.2018.01.010

39 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.

40 Ho, D., Imai, K., King, G., Stuart, E., & Whitworth, A. (2018). Package 'MatchIt'. In: Version.

41 Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw, 42*(8), 1-28.

42 Jack, C. R., Jr., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., . . . Trojanowski, J. Q. (2013). Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol, 12*(2), 207-216. doi:10.1016/S1474-4422(12)70291-0

43 Jack, C. R., Therneau, T. M., Weigand, S. D., Wiste, H. J., Knopman, D. S., Vemuri, P., . . . Machulda, M. M. (2019). Prevalence of biologically vs clinically defined Alzheimer spectrum entities using the National Institute on Aging–Alzheimer's Association research framework. *JAMA neurology, 76*(10), 1174-1183.

44 Jack Jr, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., . . . Karlawish, J. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562.

45 Jack Jr, C. R., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., Knopman, D. S., . . . Kantarci, K. (2017). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimer's & Dementia, 13*(3), 205-216.

46 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 : official journal of the American Association for Geriatric Psychiatry, 17*(5), 368-375. doi:10.1097/jgp.0b013e31819431d5

47 Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A., Jones, R. N., Choi, S. E., . . . Tommet, D. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1), e12055.

48 Jutten, R. J., Thompson, L., Sikkes, S. A., Maruff, P., Molinuevo, J. L., Zetterberg, H., . . . Gold, M. (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*(Preprint), 1-14.

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

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

51 Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., . . . 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.

52 Malek-Ahmadi, M. (2016). Reversion from mild cognitive impairment to normal cognition. *Alzheimer Disease & Associated Disorders, 30*(4), 324-330.

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

54 Mathews, M., Abner, E., Kryscio, R., Jicha, G., Cooper, G., Smith, C., . . . 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.

55 Mehta, C., Gao, P., Bhatt, D. L., Harrington, R. A., Skerjanec, S., & Ware, J. H. (2009). Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation, 119*(4), 597-605.

56 Mitchell, A. J., & Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica, 119*(4), 252-265.

57 Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., . . . Strobel, G. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology, 15*(7), 673-684.

58 Pandya, S. Y., Clem, M. A., Silva, L. M., & Woon, F. L. (2016). Does mild cognitive impairment always lead to dementia? A review. *Journal of the neurological sciences, 369*, 57-62.

59 Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., . . . Toga, A. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology, 74*(3), 201-209.

60 Qiu, W., & Qiu, M. W. (2020). Package 'powerMediation'.

61 Rafii, M. S., & Aisen, P. S. (2019). Alzheimer's Disease Clinical Trials: Moving Toward Successful Prevention. *CNS drugs, 33*(2), 99-106.

62 Rönnlund, M., & Nilsson, L.-G. (2006). Adult life-span patterns in WAIS-R Block Design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence, 34*(1), 63-78.

63 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.

64 Ronnlund, M., Nyberg, L., Backman, 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. *Psychol Aging, 20*(1), 3-18. doi:10.1037/0882-7974.20.1.3

65 Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International neuropsychological Society, 16*(5), 754-760.

66 Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and aging, 34*(1), 17.

67 Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., . . . Eppig, J. S. (2021). Cognitive Practice Effects Delay Diagnosis; Implications for Clinical Trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions, Preprint*.

68 Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. 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*. doi:10.3389/fnagi.2022.847315

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

70 Schrijnemaekers, A., de Jager, C. A., Hogervorst, E., & Budge, 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.

71 Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., . . . Alzheimer's Disease Neuroimaging, I. (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Ann Neurol, 65*(4), 403-413. doi:10.1002/ana.21610

72 Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy, 3*(6), 32.

73 Sperling, R., Mormino, E., & Johnson, K. (2014). The evolution of preclinical Alzheimer's disease: implications for prevention trials. *Neuron, 84*(3), 608-622.

74 Sperling, R. A., Donohue, M. C., Raman, R., Sun, C.-K., Yaari, R., Holdridge, K., . . . Aisen, P. S. (2020). Association of factors with elevated amyloid burden in clinically normal older individuals. *JAMA neurology*.

75 Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., & Aisen, P. (2014). The A4 study: stopping AD before symptoms begin? *Science translational medicine, 6*(228), 228fs213-228fs213.

76 Stricker, N. H., Lundt, E. S., Alden, E. C., Albertson, S. M., Machulda, M. M., Kremers, W. K., . . . 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.

77 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.

78 Team, R. C. (2019). language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

79 Thomas, K. R., Cook, S. E., Bondi, M. W., Unverzagt, F. W., Gross, A. L., Willis, S. L., & Marsiske, M. (2020). Application of neuropsychological criteria to classify mild cognitive impairment in the active study. *Neuropsychology, 34*(8), 862.

80 Thomas, K. R., Edmonds, E. C., Delano-Wood, L., & Bondi, M. W. (2017). Longitudinal trajectories of informant-reported daily functioning in empirically defined subtypes of mild cognitive impairment. *Journal of the International Neuropsychological Society, 23*(6), 521-527.

81 Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., . . . Salmon, D. P. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia, 15*(10), 1322-1332.

82 Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., . . . Morris, J. C. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia, 15*(1), 106-152.

83 Vivot, A., Power, M. C., Glymour, M. M., Mayeda, E. R., Benitez, A., Spiro III, A., . . . 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.

84 Vuoksimaa, E., McEvoy, L. K., Holland, D., Franz, C. E., & Kremen, W. S. (2020). Modifying the minimum criteria for diagnosing amnestic MCI to improve prediction of brain atrophy and progression to Alzheimer's disease. *Brain imaging and behavior, 14*(3), 787-796.

85 Wang, G., Kennedy, R. E., Goldberg, T. E., Fowler, M. E., Cutter, G. R., & Schneider, L. S. (2020). Using practice effects for targeted trials or sub-group analysis in Alzheimer's disease: How practice effects predict change over time. *PloS one, 15*(2), e0228064. doi:10.1371/journal.pone.0228064

86 Wilson, B. A., Watson, P. C., Baddeley, A. D., Hazel, E., & Evans, J. J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *Journal of the International Neuropsychological Society, 6*(4), 469-479.

87 Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., . . . Almkvist, O. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *Journal of internal medicine, 256*(3), 240-246.