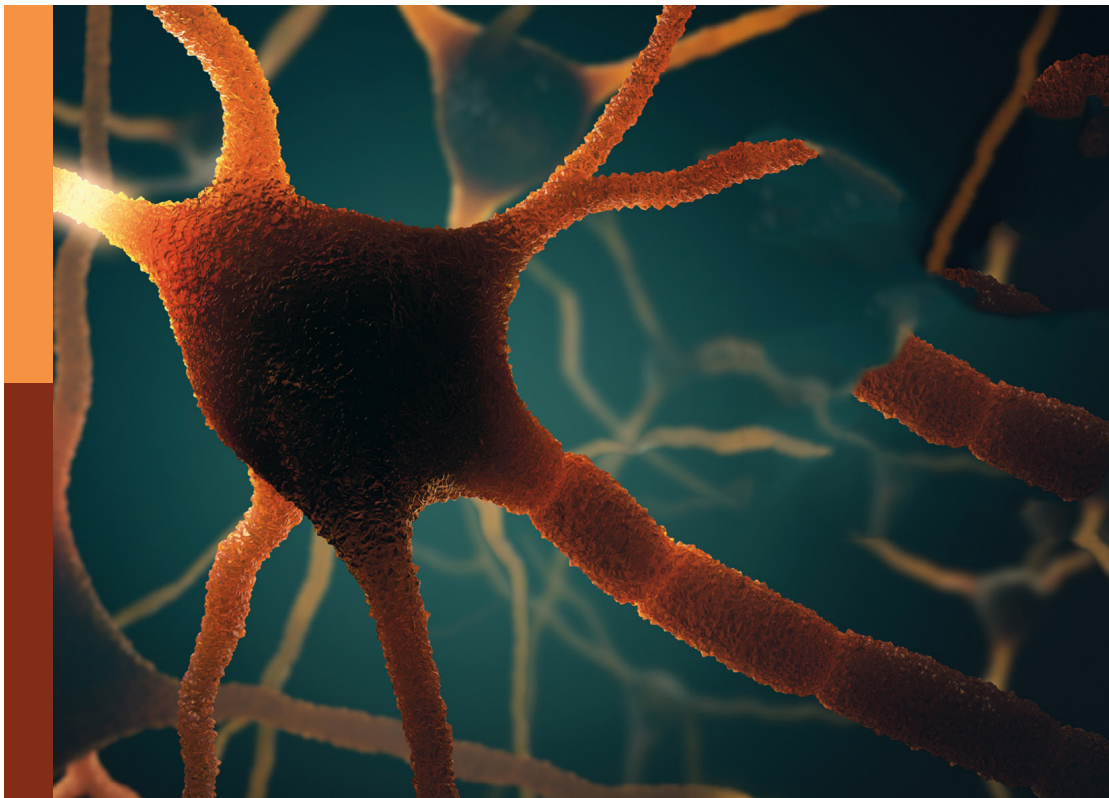
frontiers | Research Topics

The importance of cognitive practice effects in aging neuroscience

Edited by
William Kremen, Daniel Nation and Lars Nyberg

Published in
Frontiers in Aging Neuroscience



ISSN 1664-8714
ISBN 978-2-83250-939-5
DOI 10.3389/978-2-83250-939-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

The importance of cognitive practice effects in aging neuroscience

Topic editors

William Kremen — University of California, San Diego, United States
Daniel Nation — University of California, Irvine, United States
Lars Nyberg — Umeå University, Sweden

Citation

Kremen, W., Nation, D., Nyberg, L., eds. (2022). *The importance of cognitive practice effects in aging neuroscience*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-939-5

Table of contents

05 Editorial: The importance of cognitive practice effects in aging neuroscience
William S. Kremen, Daniel A. Nation and Lars Nyberg

08 Monthly At-Home Computerized Cognitive Testing to Detect Diminished Practice Effects in Preclinical Alzheimer's Disease
Roos J. Jutten, Dorene M. Rentz, Jessie F. Fu, Danielle V. Mayblyum, Rebecca E. Amariglio, Rachel F. Buckley, Michael J. Properzi, Paul Maruff, Craig E. Stark, Michael A. Yassa, Keith A. Johnson, Reisa A. Sperling and Kathryn V. Papp

21 Episodic Memory and Executive Function Are Differentially Affected by Retests but Similarly Affected by Age in a Longitudinal Study of Normally-Aging Older Adults
Elizabeth L. Glisky, Cindy B. Woolverton, Katelyn S. McVeigh and Matthew D. Grilli

34 Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments
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 for the Alzheimer's Disease Neuroimaging Initiative

48 Parameterizing Practice in a Longitudinal Measurement Burst Design to Dissociate Retest Effects From Developmental Change: Implications for Aging Neuroscience
Nicholas Tamburri, Cynthia McDowell and Stuart W. S. MacDonald

62 Avoid or Embrace? Practice Effects in Alzheimer's Disease Prevention Trials
Andrew J. Aschenbrenner, Jason Hassenstab, Guoqiao Wang, Yan Li, Chengjie Xiong, Eric McDade, David B. Clifford, Stephen Salloway, Martin Farlow, Roy Yaari, Eden Y. J. Cheng, Karen C. Holdridge, Catherine J. Mummery, Colin L. Masters, Ging-Yuek Hsiung, Ghulam Surti, Gregory S. Day, Sandra Weintraub, Lawrence S. Honig, James E. Galvin, John M. Ringman, William S. Brooks, Nick C. Fox, Peter J. Snyder, Kazushi Suzuki, Hiroyuki Shimada, Susanne Gräber and Randall J. Bateman for the Dominantly Inherited Alzheimer Network Trials Unit (DIAN-TU)

73 Practice Effect of Repeated Cognitive Tests Among Older Adults: Associations With Brain Amyloid Pathology and Other Influencing Factors
Bang Zheng, Chinedu Udeh-Momoh, Tamlyn Watermeyer, Celeste A. de Jager Loots, Jamie K. Ford, Catherine E. Robb, Parthenia Giannakopoulou, Sara Ahmadi-Abhari, Susan Baker, Gerald P. Novak, Geraint Price and Lefkos T. Middleton

81 Neuropsychological Decline Stratifies Dementia Risk in Cognitively Unimpaired and Impaired Older Adults
Jean K. Ho and Daniel A. Nation

90 Dynamic modeling of practice effects across the healthy aging-Alzheimer's disease continuum
Andrew R. Bender, Arkaprabha Ganguli, Melinda Meiring, Benjamin M. Hampstead and Charles C. Driver

107 Cognitive and structural predictors of novel task learning, and contextual predictors of time series of daily task performance during the learning period
Evan T. Smith, Paulina Skolasinska, Shuo Qin, Andrew Sun, Paul Fishwick, Denise C. Park and Chandramallika Basak

128 Accounting for retest effects in cognitive testing with the Bayesian double exponential model via intensive measurement burst designs
Zita Oravecz, Karra D. Harrington, Jonathan G. Hakun, Mindy J. Katz, Cuiling Wang, Ruixue Zhaoyang and Martin J. Sliwinski

147 Practice effects in cognitive assessments three years later in non-carriers but not in symptom-free mutation carriers of autosomal-dominant Alzheimer's disease: Exemplifying procedural learning and memory?
Ove Almkvist and Caroline Graff

frontiers | Frontiers in Aging Neuroscience

TYPE Editorial
PUBLISHED 18 November 2022
DOI 10.3389/fnagi.2022.1079021

OPEN ACCESS

EDITED AND REVIEWED BY
Kristy A. Nielson,
Marquette University, United States

*CORRESPONDENCE
William S. Kremen
wkremen@ucsd.edu

SPECIALTY SECTION
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 24 October 2022
ACCEPTED 07 November 2022
PUBLISHED 18 November 2022

CITATION
Kremen WS, Nation DA and Nyberg L (2022) Editorial: The importance of cognitive practice effects in aging neuroscience.
Front. Aging Neurosci. 14:1079021.
doi: 10.3389/fnagi.2022.1079021

Editorial: The importance of cognitive practice effects in aging neuroscience

William S. Kremen[1]*, Daniel A. Nation[2] and Lars Nyberg[3,4]

[1]Department of Psychiatry and Center for Behavior Genetics of Aging, University of California San Diego, La Jolla, CA, United States, [2]Department of Psychological Science, University of California, Irvine, Irvine, CA, United States, [3]Department of Radiation Sciences, Umeå University, Umeå, Sweden, [4]Department of Integrative Medical Biology, Umeå Center for Functional Brain Imaging, Umeå University, Umeå, Sweden

KEYWORDS

practice effects, cognitive aging, mild cognitive impairment, Alzheimer's disease, dementia progression, cognitive intervention

Editorial on the Research Topic
The importance of cognitive practice effects in aging neuroscience

Practice effects (PEs) on repeated cognitive testing is a well-known phenomenon, yet it is rarely systematically taken into account and most often simply ignored. However, failure to account for PEs can have a substantial negative impact in aging neuroscience. This Featured Resarch Topic includes 11 original research papers (cited in this editorial). We have divided them into seven non-mutually exclusive categories: (1) using level of PEs to improve prediction of progression to cognitive impairment status (Almkvist and Graff; Aschenbrenner et al.; Bender et al.; Ho and Nation; Jutten et al.; Tamburri et al.; Zheng et al.); (2) identifying predictors of reduced PEs (Bender et al.; Glisky et al.; Jutten et al.; Zheng et al.); (3) examining the magnitude of PEs associated with diagnostic severity—from cognitively unimpaired, to mild cognitive impairment (MCI) to dementia (Ho and Nation; Jutten et al.; Oravecz et al.; Tamburri et al.), or from asymptomatic mutation carriers to symptomatic mutation carriers to autosomal dominant Alzhemer's dementia (Almkvist and Graff; Aschenbrenner et al.); (4) examining PEs in normal aging (Glisky et al.); (5) adjusting cognitive scores for PEs to detect MCI earlier and characterize its progression more accurately (Sanderson-Cimino et al.); (6) using burst designs and dynamic modeling to differentiate short-term and long-term PE fluctuations and to focus on intraindividual variability (Bender et al.; Oravecz et al.; Tamburri et al.); and (7) using PEs to improve evaluation of cognitive interventions (Smith et al.).

On the surface, PEs seem simple and straightforward, i.e., they are improvements in performance on repeated testing. However, lack of improvement, and even cognitive decline, does not necessarily mean an absence of PEs. As aptly noted by some authors, it may only mean that normal aging-related or disease-related declines were still greater than the PEs (Aschenbrenner et al.; Glisky et al.; Sanderson-Cimino et al.).

All too often, we find that people are interested in which is the best method for examining PEs, frequently wanting to know if the approach being used is as good as

some other approach or suggesting another approach would be preferable. Importantly, here we want to emphasize that different approaches often address entirely different issues and serve very different purposes, so trying to determine which is best is often a misguided goal. There is simply no one-size-fits-all approach. For example, several of the articles addressed the issue described in category 1 above in which the extent of PEs was used to predict individuals who would likely progress to MCI, Alzheimer's disease (AD), or other dementia (Aschenbrenner et al.; Bender et al.; Ho and Nation; Jutten et al.; Tamburri et al.; Zheng et al.). Addressing the issue described in category 5 above, Sanderson-Cimino et al. adjusted test scores for PEs based on comparison of test-naïve vs. returning participants. Doing so meant that MCI could be detected earlier and MCI progression characterized more accurately. Both sets of methods provide useful adjunctive tools for improving clinical trials and diagnostic accuracy, yet one is in no way substitutable for the other. The former approach does nothing to alter how or when the diagnosis is made. The latter does nothing to aid in predicting progression to diagnosis.

Here we note some key take-home messages regarding PEs:

1. Some studies define PEs as improvement in performance on retesting (Almkvist and Graff) or as improvement on short-term, but not long-term, retest intervals (Oravecz et al.; Tamburri et al.). However PEs are also consistently observed over intervals of a year or more (Almkvist and Graff; Bender et al.; Glisky et al.; Sanderson-Cimino et al.). Therefore, we suggest that improvements or reduced declines be referred to as PEs regardless of the size of the test-retest interval.
2. PEs make it difficult to disentangle aging-related and disease-related effects. Thus, PEs mask normal aging-related cognitive change, making it difficult to accurately characterize the course of longitudinal change. Only with matched previously untested participants at follow-up is it possible to accurately distinguish among change, effects of attrition, and PEs.
3. There is no general cognitive PE, which raises questions about the usefulness of global cognitive measures to assess PEs. It should not be assumed that the magnitude of PEs from one study would apply to another study. PEs may differ depending on:
    a. Cognitive domain
    b. Tests within a domain
    c. Age
    d. Diagnosis
    e. Duration of test-retest interval
    f. Number of repeat assessments
    g. Risk factors (e.g., AD biomarker status, brain structure, sleep, psychological wellbeing)
4. Alternate forms have been suggested as a possible way to reduce PEs (Aschenbrenner et al.). However, alternate forms make it more difficult to differentiate actual PEs from test version differences.
5. Slope of change (extent of PEs) may be a better predictor of progression to diagnosis than baseline level of function (Jutten et al.).
6. Burst designs or monthly testing are effective ways to characterize change and can be particularly useful for improved understanding of the dynamics of cognitive change, and they highlight the additional potential predictive value of within-individual variability in PEs (Bender et al.; Jutten et al.; Oravecz et al.; Tamburri et al.).
7. PEs can be usefully applied in cognitive interventions for prediction of likelihood of benefit and of transfer of training (Smith et al.).
8. Accounting for PEs by comparisons with matched previously untested participants at follow-up, results in earlier and more accurate diagnosis based on associations with reduced reversion rates of MCI and greater concordance with AD biomarkers (Sanderson-Cimino et al.).

In sum, accounting for cognitive PEs is important for accurately characterizing longitudinal change and progression to cognitive impairment status, and it is crucial to do it in a way that differentiates PEs from aging-related or disease-related change. Given the many factors that influence PEs, the magnitude of PEs cannot be expected to be comparable across studies. Incorporating PEs into clinical trials can improve participant selection efficiency and result in earlier detection of diagnostic outcomes. Such changes could also reduce study duration and staff and participant burden, which in turn, would substantially reduce costs. Only a single study in this set of papers examined PEs in the context of a cognitive intervention. Also, only a single study included matched previously untested participants at follow-up. Such matched replacements are critical for accurately distinguishing among change, the effects of attrition, and PEs. Although normative data might appear to be a solution, it provides no insight into the actual magnitude of PEs for a given age group. Given that the goals of these latter 2 studies are of great potential value, more work is called for in these areas in addition to the other areas of focus in research on cognitive PEs.

Author contributions

WK, DN, and LN contributed to the conception and interpretation of results described in this editorial. All authors contributed to the article and approved the submitted version.

Funding

WK was supported by grants from the United States National Institute on Aging (NIA; R01 AG050595, R01 AG876838, AG037985, AG062483, AG064955, and P01 AG055367). DN was supported by grants from the NIA (R01 AG064228, R01 AG060049, P01 AG052350, and P30 AG066519). LN was supported by a scholar grant from KAW.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fnagi.2021.800126

Monthly At-Home Computerized Cognitive Testing to Detect Diminished Practice Effects in Preclinical Alzheimer's Disease

*Roos J. Jutten[1], Dorene M. Rentz[1,2], Jessie F. Fu[3], Danielle V. Mayblyum[3], Rebecca E. Amariglio[1,2], Rachel F. Buckley[1,4], Michael J. Properzi[1], Paul Maruff[5,6], Craig E. Stark[7], Michael A. Yassa[7], Keith A. Johnson[1,3], Reisa A. Sperling[1,2] and Kathryn V. Papp[1,2]**

[1] Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, [2] Department of Neurology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, [3] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States, [4] Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia, [5] CogState Ltd., Melbourne, VIC, Australia, [6] The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Melbourne, VIC, Australia, [7] Department of Neurobiology and Behavior, Center for the Neurobiology of Learning and Memory, University of California, Irvine, Irvine, CA, United States

OPEN ACCESS

Edited by:
William Kremen, University of California, San Diego, United States

Reviewed by:
Hillary Protas, Banner Alzheimer's Institute, United States
Rosaleena Mohanty, Karolinska Institutet (KI), Sweden

***Correspondence:**
Kathryn V. Papp
kpapp@bwh.harvard.edu

Specialty section:
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

Received: *22 October 2021*
Accepted: *14 December 2021*
Published: *13 January 2022*

Citation:
Jutten RJ, Rentz DM, Fu JF, Mayblyum DV, Amariglio RE, Buckley RF, Properzi MJ, Maruff P, Stark CE, Yassa MA, Johnson KA, Sperling RA and Papp KV (2022) Monthly At-Home Computerized Cognitive Testing to Detect Diminished Practice Effects in Preclinical Alzheimer's Disease. Front. Aging Neurosci. 13:800126. doi: 10.3389/fnagi.2021.800126

Introduction: We investigated whether monthly assessments of a computerized cognitive composite (C3) could aid in the detection of differences in practice effects (PE) in clinically unimpaired (CU) older adults, and whether diminished PE were associated with Alzheimer's disease (AD) biomarkers and annual cognitive decline.

Materials and Methods: $N = 114$ CU participants (age 77.6 ± 5.0, 61% female, MMSE 29 ± 1.2) from the Harvard Aging Brain Study completed the self-administered C3 monthly, at-home, on an iPad for one year. At baseline, participants underwent in-clinic Preclinical Alzheimer's Cognitive Composite-5 (PACC5) testing, and a subsample ($n = 72$, age = 77.8 ± 4.9, 59% female, MMSE 29 ± 1.3) had 1-year follow-up in-clinic PACC5 testing available. Participants had undergone PIB-PET imaging (0.99 ± 1.6 years before at-home baseline) and Flortaucipir PET imaging ($n = 105$, 0.62 ± 1.1 years before at-home baseline). Linear mixed models were used to investigate change over months on the C3 adjusting for age, sex, and years of education, and to extract individual covariate-adjusted slopes over the first 3 months. We investigated the association of 3-month C3 slopes with global amyloid burden and tau deposition in eight predefined regions of interest, and conducted Receiver Operating Characteristic analyses to examine how accurately 3-month C3 slopes could identify individuals that showed >0.10 SD annual decline on the PACC-5.

Results: Overall, individuals improved on all C3 measures over 12 months ($\beta = 0.23$, 95% CI [0.21–0.25], $p < 0.001$), but improvement over the first 3 months was greatest ($\beta = 0.68$, 95% CI [0.59–0.77], $p < 0.001$), suggesting stronger PE over initial repeated exposures. However, lower PE over 3 months were associated with more global amyloid burden ($r = -0.20$, 95% CI [−0.38 – −0.01], $p = 0.049$) and tau deposition in the entorhinal cortex ($r = -0.38$, 95% CI [−0.54 – −0.19], $p < 0.001$) and inferior-temporal

lobe ($r = -0.23$, 95% CI [−0.41 – −0.02], $p = 0.03$). 3-month C3 slopes exhibited good discriminative ability to identify PACC-5 decliners (AUC 0.91, 95% CI [0.84–0.98]), which was better than baseline C3 ($p < 0.001$) and baseline PACC-5 scores ($p = 0.02$).

Conclusion: While PE are commonly observed among CU adults, diminished PE over monthly cognitive testing are associated with greater AD biomarker burden and cognitive decline. Our findings imply that unsupervised computerized testing using monthly retest paradigms can provide rapid detection of diminished PE indicative of future cognitive decline in preclinical AD.

Keywords: computerized testing, remote assessment, practice effects, digital biomarkers, preclinical AD

INTRODUCTION

Alongside the increased focus on characterizing Alzheimer's disease (AD) in the preclinical stage, there is a need to detect and track the cognitive changes that may emerge during this stage more rapidly. However, capturing short-term cognitive changes in preclinical AD is a major challenge using conventional paper-and-pencil cognitive tests, which typically require in-clinic assessments at annual intervals and only detect subtle decline over multiple years (Petersen et al., 2016; Mortamais et al., 2017; Jutten et al., 2020b). This is a particular hurdle for AD secondary prevention trials, which currently require large sample-sizes and lengthy follow-up to enable the detection of an attenuation of subtle cognitive decline.

Computerized cognitive testing has the potential to capture changes in cognition earlier, by enabling standardized administration and data analyses allowing for remote, unsupervised, and more frequent assessments (e.g., monthly rather than yearly) in a feasible way (Gold et al., 2018; Koo and Vizer, 2019). Several computerized tests have been developed for use in remote, unsupervised settings, including the Computerized Cognitive Composite (C3) battery, which was designed to assess cognitive processes that rely on the medial temporal lobe (MTL) (Rentz et al., 2016; Buckley et al., 2017; Papp et al., 2021b). The C3 comprises two well-validated episodic memory paradigms: the Face Name Associative Memory Exam (FNAME) (Rentz et al., 2011) and the Behavioral Pattern Separation Task—Object Version (BPSO) (Stark et al., 2013), and the Cogstate Brief Battery (CBB) (Maruff et al., 2009; Lim et al., 2012). It was recently shown that unsupervised, at-home C3 testing on an iPad was feasible and could provide data that discriminated reliably between cognitively normal and impaired adults (Rentz et al., 2016; Buckley et al., 2017; Papp et al., 2021b).

The higher frequency assessments afforded through use of computerized tests enable the study of practice effects (PE) that can occur with repeated cognitive assessments (Beglinger et al., 2005). PE have typically been viewed as a source of bias (Salthouse, 2012), but several studies showed that characterizing PE could provide an indicator of cognitive impairment and, more specifically, that lower PE reflect a decreased ability to benefit from previous experience when re-exposed ot the same stimuli (Duff et al., 2007, 2012; Jutten et al., 2020a). PE have been reported for the individual C3 and CBB measures when administered in clinically unimpaired (CU) adults (Baker et al., 2019; Samaroo et al., 2020; Stricker et al., 2020). Interestingly, the study by Samaroo et al. revealed *diminished* PE on the FNAME test in CU with high levels of amyloid compared to CU with low levels of amyloid, which was evident from only 4 months of repeated assessments. This suggests that failure of learning due to practice may already be evident in preclinical AD, and that the magnitude of PE may have potential as a cognitive marker of this very early manifestation of the disease.

The current study expands on previous work by investigating whether characterizing PE across a range of memory tasks included in the C3 battery could aid in the detection of early cognitive change in preclinical AD. First, we investigated the nature and magnitude of PE that arose from monthly repeated exposure to at-home C3 assessments over 1 year. Upon seeing improvement, we investigated whether PE on computerized testing could be observed over the first 3 months, as we expected that the PE signal would be strongest over the first 4–5 assessments (Watson et al., 1994; Calamia et al., 2012; Samaroo et al., 2020). Next, we examined the relationship of individual variation in shorter term PE (i.e., 3 months) with (1) AD biomarker burden measured using neuroimaging and (2) cognitive decline on standard paper-pencil cognitive testing over 1 year (Petersen et al., 2021).

MATERIALS AND METHODS

Study Participants

The current study describes data from the At-Home Digital Cognition Sub-Study including participants from the Harvard Aging Brain Study (HABS). HABS is an ongoing longitudinal observational cohort-study of community-dwelling older adults who are clinically normal at the time of enrollment. Inclusion criteria for HABS have been described in detail elsewhere (Dagley et al., 2017). The At-Home Digital Cognition Study started recruiting participants in the 6th HABS year. For participation in the At-Home Digital Cognition Study, participants were deemed to be CU at the start of the study, which was determined by clinician consensus based on cognitive and functional test results and medical history (Papp et al., 2020). The study was approved by an ethical review board, and all participants provided written informed consent.

Measures

Computerized Cognitive Composite

The C3 battery is a self-administered test battery presented on an iPad using CogState software. It includes the Face Name Associative Memory Exam (FNAME) (Rentz et al., 2011), a version of the Behavioral Pattern Separation Task-Object version (BPSO) (Stark et al., 2013), and the Cogstate Brief Battery comprising four brief tests: the Detection Task (DET), the Identification Task (IDN), One Card Learning Task (OCL) and the One-Back Task (ONB) (Maruff et al., 2009). The C3 and its individual measures have been described in detail elsewhere (Papp et al., 2021b), and **Supplementary Table 1** provides a detailed overview of the individual outcomes.

Briefly, the FNAME is an associative memory paradigm requiring participants to encode and subsequently recall and match faces with corresponding names. Participants are shown 12 face-name pairs, and after a 12–15-min delay there are three measures of memory including first letter name recall, face-name matching and face recognition. For the current study, we focused on the free recall measure, i.e., the first letter name recall test (FNLT), since this is the FNAME measure that is expected to have the fewest range restrictions in scoring and therefore most likely to capture PE over repeated exposures. Participants are asked to select the first letter of the name paired with that face, and the primary outcome is the number of first letters correctly recalled. Total score range is 0–12 with higher scores reflecting better performance.

For the BPSO, participants are presented with a series of unique images (encoding phase) and encouraged to attend carefully to the physical characteristics of each object by having them decide whether the object is used mostly outdoors or indoors. This is followed by a recognition phase that includes repeated, novel and distractor images (lures), which participants are asked to categorize into Old, Similar, or New. Of the images presented during the recognition phase, one third are identical to those presented during encoding (for which the correct response would be "Old"), one third of the images contains an object that is visually similar, but not identical to an object presented during the encoding phase (i.e., lures, for which correct response would be "Similar") and one third are objects that had not been seen during encoding (i.e., foils, for which the correct response would be "New"). The version of the BPSO that was used in the current study differs from the original version in that the studied items brought into the recognition phase are presented both as repeated identical targets and as similar lures, with half of the items having the target version presented first and half of the items having the similar lure version presented first. The primary outcome of the BPSO is a metric reflecting the ability to correctly discriminate between stimuli that are similar but not identical to previously learned items. That is, a Lure Discrimination Index (LDI) is calculated as: Proportion of "similar" responses made to Lure trials minus the proportion of "similar" responses to Foil trials. The LDI range is 0–1, with higher scores reflecting better performance.

The CBB uses playing cards as stimuli to measure reaction time and working memory. The DET is a measure of attention, and participants are asked to respond when a stimulus card is turned face up. The IDN is a measure of attention and inhibitory control, where a respondent must choose whether a flipped card is red or not. Primary outcome measures for the DET and IDN are reaction time. The OCL task is a non-verbal continuous memory task in which playing cards are shown one at a time with a subset of the cards repeating several times throughout the task. The ONB task measures working memory by requiring participants to serially match each card to the previous trial. Outcome measures for the OCL and ONB include both reaction time and number of correct responses.

In-clinic Cognitive Testing

Participants underwent standard paper-and-pencil in-clinic cognitive testing including the Preclinical Alzheimer's Cognitive Composite 5 (PACC-5) (Donohue et al., 2014; Papp et al., 2017). The PACC5 is a widely used cognitive outcome measure in research and clinical trials of preclinical AD and comprises well-validated paper-and-pencil tests including the Mini-Mental State Examination (MMSE) (Folstein et al., 1975), the Wechsler Memory Scale-Revised Logical Memory Delayed Recall (Wechsler, 1987), the Digit-Symbol Coding Test (Wechsler, 2008), the Free and Cued Selective Reminding Test Free + Total Recall (Grober et al., 2009), and the Category Fluency Test (Monsch et al., 1992). Here, the PACC5 is computed as an averaged z-score of all individual measures.

Amyloid and Tau Biomarkers

We used neuro imaging to investigate whether the magnitude of PE was associated with global amyloid burden and regional tau deposition, since our current understanding of preclinical AD is that amyloid pathology is diffusely distributed across brain areas (Villemagne et al., 2011; Mormino et al., 2014) whereas tau deposition is initially focally present in the MTL regions (Johnson et al., 2016; Hanseeuw et al., 2019) where it is found to be associated with episodic performance (Maass et al., 2018). Amyloid burden and tau deposition were measured and quantified using positron-emission tomography (PET) imaging using ^{11}C-Pittsburg Compound-B (PiB) and ^{18}F-Flortaucipir (FTP), respectively, in accordance with established protocols for acquisition and analysis (Mormino et al., 2014; Johnson et al., 2016). Briefly, PiB images were acquired using a 60-min dynamic acquisition and FTP images were acquired from 75 to 105 min post-injection on a Siemens ECAT HR+ PET scanner. Following acquisition, a mean PET image was created and coregistered with the corresponding T1 MR image using the SPM12 package (Wellcome Centre for Human Neuroimaing) and the resulting coregistration transformation matrices were saved. FreeSurfer (v6) regions of interest (ROIs) defined by segmenting the MR images were transformed into the PET native space using the inverse transformation matrices. PiB was expressed as the distribution volume ratio (DVR, estimated with reference Logan graphical method), and FTP as an averaged standardized uptake value ratio (SUVR) over 70–105 min corrected for partial volume effects (PVC). For both PiB and FTP, bilateral cerebellum gray matter was used as the reference region for DVR and SUVr estimates respectively. For PiB, a global cortical aggregate was calculated for each participant based on the average PiB DVR in

frontal, lateral temporoparietal, and retrosplenial (FLR) regions, and participants were dichotomized into low (Aβ−) vs. high (Aβ+) groups (DVR cut-off-1.185). For FTP, we used the SUVR PVC values of eight predefined ROIs: the entorhinal cortex, inferior temporal lobe, amygdala, hippocampus [adjusted for choroid plexus (Lee et al., 2018)], parahippocampal region, fusiform gyrus, precuneus and posterior cingulate region.

Procedures

Baseline and conclusion of the At-Home Digital Cognition Study coincided with participants' annual HABS in-clinic visits. At baseline (In Clinic Visit 1), participants completed an iPad/Cogstate one-on-one training session with a trained HABS rater and completed the first C3 assessment in the clinic. Participants were then provided a study iPad to complete the C3 at home. The first At-Home C3 assessment was done independently at-home 1 week later (hereafter referred to as visit 0.25). Thereafter, participants completed the monthly C3 for 12 At-Home sessions with 4-week intervals. The final C3 administration occurred in-clinic as part of the second annual HABS visit (In-Clinic Visit 2), leading to a maximum of 15 C3 sessions. Participants received reminder calls prior to their scheduled test dates and were encouraged to complete the C3 at the same time monthly.

The C3 battery has a total administration time of 25–30 min. On screen instructions are provided, but participants do not receive feedback upon completion of any of the individual tests nor monthly assessments. Previous work indicated good feasibility and usability in unsupervised settings after one in-clinic training session, with a high percentage of older individuals completed at-home assessments correctly including those with lower computer literacy (Rentz et al., 2016; Samaroo et al., 2020; Papp et al., 2021a).

The At-Home Digital Cognition study was initially designed with all C3 tests being repeated using alternating versions. However, a second version of the FNAME was added as well, repeating the same version, based on the hypothesis that repeating the same versions would lead to stronger PE. A recent study comparing monthly performance on the FNAME alternate vs. same versions confirmed this (Samaroo et al., 2020), and we therefore decided to focus on the FNAME same version in the current study. Thus, retest procedures differed across individual C3 measures investigated in the current study. For the FNAME, the same version was repeated each month (A-A-A-A). For the BPSO, four alternate versions were used following the same sequence for everyone (A-B-C-D). For the CBB measures, alternate versions were used each month and the sequence of versions was randomized for each participant.

Statistical Analyses

Prior to statistical analyses, completion and performance checks were performed on all individual C3 measures to ensure the integrity of the data, by applying previously defined task-specific cut-offs (**Supplementary Table 1**). Scores that fell below these cut-offs were excluded from further analyses.

Statistical analyses were conducted in R version 4.0.3. Statistical significance was set at $p < 0.05$. To facilitate comparison across C3 measures, all data from all individual C3 measures were z-transformed using the overall sample mean and standard deviation (SD) at baseline. The BPSO, FNLT, and OCL accuracy z-scores were summed into an overall C3 z-score (Papp et al., 2021b). Linear mixed models (LMM) were used to investigate C3 performance over time (months, continuous) correcting for age, sex, and years of education. Since there were no significant interactions between time and covariates (i.e., age, sex, and years of education) for any of the C3 measures, interaction terms were not included in the final models. We initially ran the LMM including all follow-up data to describe monthly performance over 1 year, and subsequently repeated the same models including only follow-up data over the first 3 months to investigate the magnitude of PE over the initial assessments. Mean to standard deviation ratios (MSDRs) were calculated for each measure to compare effect-sizes across measures over 3 months. Figures showing the mean, SD and 95% confidence interval (CI) by study visit (i.e., time as categorical variable) are provided to visualize the overall trajectory of C3 performance.

Next, individual covariate-adjusted slopes were extracted from the aforementioned LMM to quantify PE over 3 months for each participant. Pearson's correlations were computed to investigate the association between 3-month C3 slopes and baseline amyloid burden (PiB DVR, continuous) as well as tau deposition in the entorhinal cortex and inferior-temporal lobe (SUVR, partial-volume corrected). After observing that correlations between C3 slopes and FTP uptake in the entorhinal and inferior temporal regions were significant, we sought to explore the relationship with a potential pattern of tau uptake in these and other regions which have shown early accumulation (Johnson et al., 2016). To that end, FTP data was analyzed using Partial Least Squares (PLS) analysis performed using MATLAB. PLS is a data reduction technique that produces predictive models when data are highly collinear, and hence it can be applied to imaging data as multivariate analysis method for identifying spatial patterns that are optimally associated with task performance (McIntosh et al., 1996). An additional advantage is that PLS analysis may be more robust to noise in the data than univariate analysis. Here, we used PLS analysis as a *post-hoc* hypothesis-driven method to complement the univariate correlational analyses. We explored associations between C3 baseline as well as C3 slope measures and spatial distributions of tau uptake across eight ROI: the entorhinal cortex, inferior temporal lobe, amygdala, hippocampus [adjusted for choroid plexus (Lee et al., 2018)], parahippocampal region, fusiform, precuneus and posterior cingulate. PLS analysis was used to decompose the input data (FTP data for the eight ROI: SUVR, all PVC) into components that are maximally correlated with C3 slopes using MATLAB build-in function "plsregress." The number of components was predefined to seven, as seven components accounted for at least 95% of the total variance in the input data based on principal component analysis (PCA, MATLAB build-in function "pca"). Only the first PLS component resulted in significant correlations between PLS scores for FTP SUVR data and C3 slopes. This remained the same when using fewer components. Therefore, the first PLS component, representing the spatial patterns of tau

uptake that most correlated with the C3 measures, was used for further interpretations. We first ran PLS analyses in the overall sample (corrected for the total PiB FLR load) separately for all C3 measures (baseline scores as well as slopes), and then repeated the analyses separately in the Aβ− and Aβ+ groups. To protect from Type I Error, a Bonferroni correction was conducted (adjusted p-value <0.005). PLS weights (representing the contribution of each ROI to the overall spatial pattern) were z-transformed, and regions with a z-score weights > 1 or < -1 were considered significant. Five-fold cross-validation was used to minimize mean square errors. Figures including the optimal spatial pattern as well as the correlation with the presentation of this pattern and C3 slopes are provided.

Finally, individual PACC5 slopes were extracted using LMM correcting for age, sex, and education for participants with baseline and 1-year follow-up PACC5 testing available ($n = 72$). Pearson's correlations were used to assess the association between 3-month slopes on the C3 and change on the PACC5 over 1 year. We then conducted Receiver Operating Characteristic (ROC) analyses to quantify how accurately C3 slopes could identify individuals who would show more than 0.10 SD decline on the PACC5 over 1 year, which has previously been suggested as a clinically meaningful cut-off for annual decline in amyloid positive cognitively normal individuals (Papp et al., 2020; Petersen et al., 2021).

TABLE 1 | Baseline characteristics for the overall sample and subsample with in-clinic follow-up after 1 year.

	Total sample ($N = 114$)	Sample with in-clinic follow-up ($n = 72$)
FU C3 assessments, M (SD) [range]	11.7 (3.2), [2–15]	12.8 (1.8), [2–15]*
N month 0.25/1/2/3	101/104/106/104	64/68/70/69
Age, M (SD)	77.6 (5.0)	77.8 (4.9)
Female, n (%)	70 (67.3%)	42 (60%)
Years of Education, M (SD)	16.5 (2.7)	16.3 (2.8)
Global CDR, 0/0.50	105/9	66/6
MMSE score, M (SD)	29.1 (1.3)	29.2 (1.2)
PACC5 score, M (SD)	0.22 (0.76)	0.29 (0.73)
PiB-PET years since C3 baseline	−0.99 ± 1.6	−0.68 ± 1.7
Global cortical amyloid (DVR)	1.21 ± 0.23	1.22 ± 0.25
Aβ status	81 Aβ−/33 Aβ+	50 Aβ−/22 Aβ+
N	105	66
FTP-PET years since C3 baseline	−0.62 ± 1.1	−0.34 ± 1.2
FTP-PET ET Tau (SUVR, PVC)	1.38 ± 0.28	1.39 ± 0.27
FTP-PET IT Tau (SUVR, PVC)	1.50 ± 0.18	1.50 ± 0.16

*N.B. *p < 0.001.*
C3, Computerized Cognitive Composite; CDR, Clinical Dementia Rating scale; MMSE, Mini-Mental State Examination; PACC-5, Preclinical Alzheimer's Cognitive Composite-−5; PET, Positron-emission tomography; PiB, ^{11}C-Pittsburg Compound-B; Aβ, Amyloid-beta; DVR, distribution volume ratio; FTP, ^{18}F-Flortaucipir; SUVR, Standardized uptake value ratio; PVC, Partial volume corrected.

RESULTS

Sample Characteristics

Baseline characteristics of the total sample ($N = 114$) as well as subsample with 1-year in-clinic follow-up (FU) available ($n = 72$) are presented in **Table 1**. All participants had undergone PiB-PET imaging (0.99 ± 1.6 years before at-home baseline) and FTP PET was available for the majority ($n = 105$, 0.62 ± 1.1 years before at-home baseline). Overall, adherence was high with an average of 11.7 (SD = 3.2) FU C3 assessments, 96% of the participants having at least 3 completed FU assessments, 91% having at least 6 completed FU assessments and 75% having completed 12 or 13 FU assessments (**Supplementary Figure 1**). Within-testing session discontinuation rates were low (3% in total over all observations on all C3 measures across all visits). Documented reasons for non-completion mainly included technological issues or lack of time. For completed assessments, integrity checks of individual assessments were high, the criterion for performance validity was met at a 99.2% for the BPSO, 99.1% for the DET, 98.9% IDN, 99% for ONB, 98.4% OCL, 99.8% for the FNLT. Compared to the total sample, the subsample ($n = 72$) with 1 year in-clinic follow-up had completed more C3 assessments ($p < 0.001$) but did not differ regarding other baseline clinical and demographic characteristics (all $p > 0.05$).

Change Over Time on Monthly C3 Assessments

Table 2 displays the time (in months) estimates obtained from LMM correcting for age, sex, and years of education for the C3 score as well as the individual C3 measures. Overall, individuals improved over 1 year on the C3 ($\beta = 0.23$, 95% CI [0.21–0.25], $p < 0.001$). However, improvement was greatest over the first 3 months ($\beta = 0.68$, 95% CI [0.59–0.77], $p < 0.001$) suggesting stronger practice over the initial exposures, which is also visualized by the mean trajectory of C3 performance over months (**Figure 1**).

When looking at the individual measures, a statistically significant improvement was observed on most individual measures over 1 year (all p-values <0.001), except for IDN reaction time (**Table 2**). Time estimates from the models including only the first 3 months were all greater than time estimates over 1 year, particularly for the BPSO and FNLT (**Table 2**). When comparing change over 3 months across the C3 measures, improvement was greater for the FNLT (MSDR 1.37) and BPSO (MSDR 0.71) as compared to the OCL and ONB accuracy measures (MSDRs 0.25 and 0.31 respectively). For both the OCL and ONB, the reaction time measures exhibited larger effect-sizes (MSDR 0.55 and 0.67 respectively) than the accuracy measures.

Diminished Practice Over 3 Months Is Associated With AD Biomarker Burden

We found moderate negative correlations between 3-month C3 slopes (covariate adjusted) and cross-sectional global amyloid burden ($r = -0.20$, 95% CI [−0.38 – −0.01], $p = 0.049$) (**Figure 2A**) as well as tau deposition in the entorhinal cortex ($r = -0.38$, 95% CI [−0.54 – −0.19], $p < 0.001$) (**Figure 2B**) and

TABLE 2 | Time estimates extracted from linear mixed models corrected for age, sex, and education.

	Monthly change over 1 year			Monthly change over first 3 months			
	Time	**95% CI**	***P*-Value**	**Time**	**95% CI**	***P*-Value**	**MSDR**
C3	0.226	0.207–0.245	<0.001	0.678	0.587–0.768	<0.001	1.39
BPSO	0.073	0.061–0.085	<0.001	0.212	0.157–0.268	<0.001	0.71
FNLT	0.098	0.088–0.108	<0.001	0.379	0.328–0.429	<0.001	1.37
OCL acc	0.051	0.041–0.060	<0.001	0.072	0.020–0.125	0.007	0.25
ONB acc	0.034	0.023–0.044	<0.001	0.100	0.041–0.158	0.001	0.31
DET rt	−0.024	−0.036 to −0.012	<0.001	−0.051	−0.103–0.000	0.052	0.18
IDN rt	−0.009	−0.020–0.001	0.073	−0.062	−0.115 to −0.010	0.021	0.22
OCL rt	−0.032	−0.044 to −0.021	<0.001	−0.141	−0.188 to −0.095	<0.001	0.55
ONB rt	−0.052	−0.063 to −0.041	<0.001	−0.163	−0.207 to −0.118	<0.001	0.67

N.B. C3 is summed z-score of BPSO + FNLT + OCL. Negative scores on reaction time measures reflect improvement.
C3, Computerized Cognitive Composite (computed as the sum of the BPSO, FNLT, and OCL accuracy z-scores); BPSO, Behavioral Pattern Separation Task—Object Version; FNLT, First Name Letter Test; OCL, One-Card Learning; ONB, One Back; DET, Detection; IDN, Identification; acc, accuracy; rt, reaction time; MSDR, mean to standard deviation ratio.

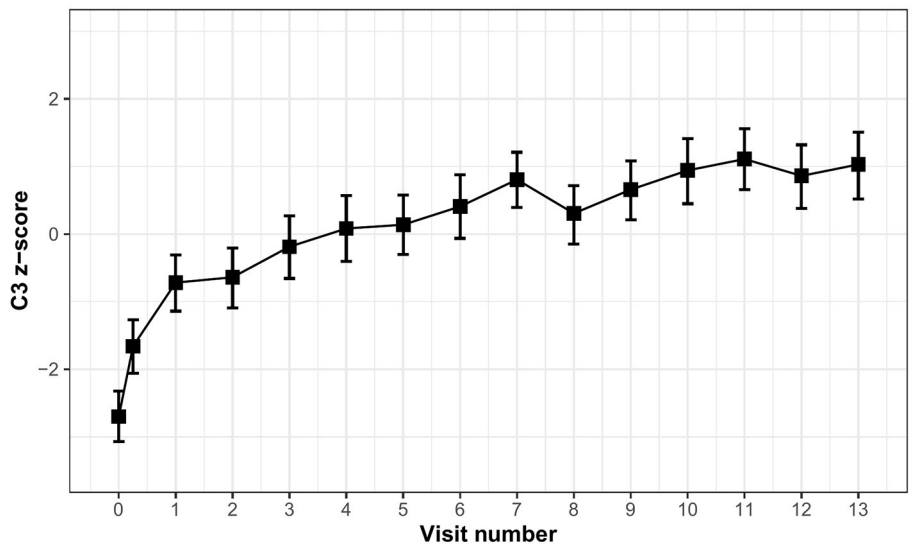


FIGURE 1 | Mean trajectory of C3 performance over monthly visits.

inferior-temporal lobe ($r = -0.23$, 95% CI [−0.41 – −0.02], $p = 0.033$) (**Figure 2C**), indicating that less improvement over 3 months is associated with greater amyloid and tau burden.

Multivariate PLS analyses revealed no spatial patterns of tau distribution that were associated with any of the C3 baseline scores in the overall sample nor in the different Aβ-groups. No spatial patterns of tau distribution were identified that significantly correlated with C3 slopes in the overall sample, but in the Aβ− group we observed a consistent spatial pattern characterized by relatively lower tau uptake in the entorhinal cortex. The expression of this spatial pattern in Aβ− group was significantly associated with higher 3-month slopes on the C3 composite ($p < 0.001$) (**Figure 3**). The correlation was most pronounced on the BPSO ($p < 0.001$) and OCL accuracy measures ($p = 0.004$) (**Figures 4**, **5**). In the Aβ+ group, we only observed a spatial pattern characterized by relatively lower tau uptake in the amygdala and entorhinal and relatively higher tau uptake in the posterior cingulate. The expression of this pattern was associated with higher 3-month slopes on the FNLT ($p = 0.003$) (**Figure 6**).

Diminished Practice Over 3 Months Is Associated With Annual Decline on the PACC5

3-month C3 slopes were positively associated with annual change on the PACC5 ($r = 0.69$, 95% CI [0.55–0.80], $p < 0.001$), indicating that less improvement over 3 months is associated with greater annual PACC5 decline (**Figure 7**).

The ROC analyses presented in **Figure 8** show that the 3-month C3 slopes exhibited good discriminative ability to identify individuals who showed >0.10 SD annual decline on the PACC5 (optimal cut-off: 0.7, area under the curve (AUC):

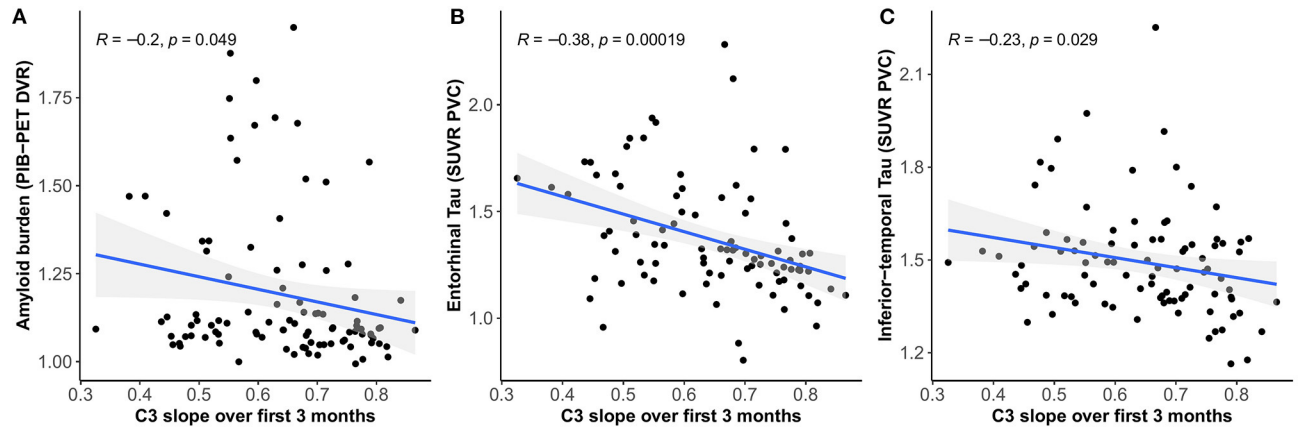


FIGURE 2 | Relationship between C3 slopes over 3 months and global amyloid burden **(A)**, tau deposition in the entorhinal cortex **(B)**, and inferior-temporal lobe **(C)**.

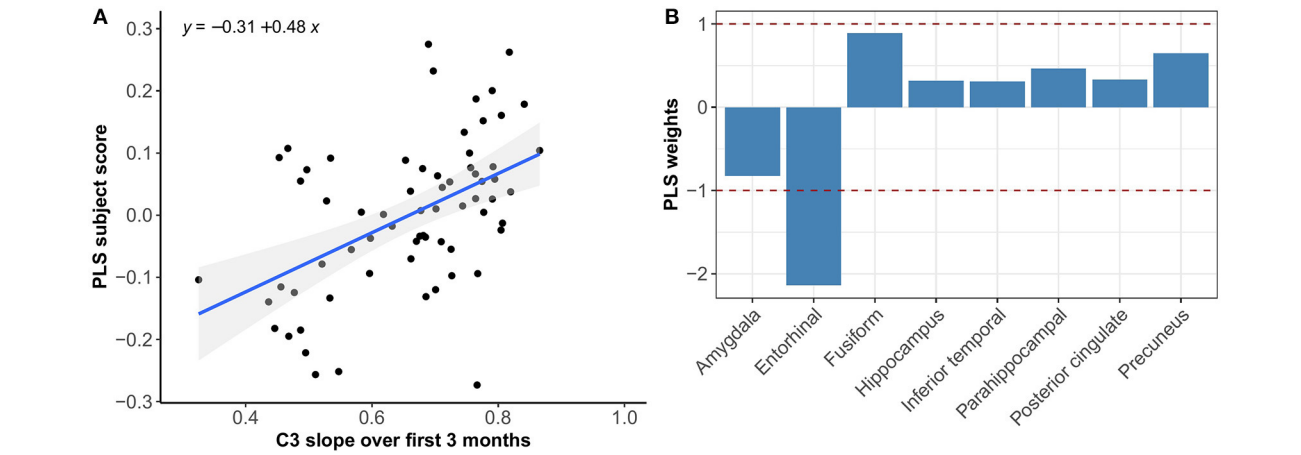


FIGURE 3 | Association between optimal spatial distribution of tau and C3 slopes over 3 months in the Aβ− group. Left panel **(A)** shows the correlation between PLS subject scores (expression of spatial patterns of tau distribution) and C3 slopes; right panel **(B)** presents the spatial patterns of tau distribution of z-transformed PLS weights in each ROI.

0.91, 95% CI [0.84–0.98], sensitivity = 88.9%, specificity = 81.1%), which was found to perform better than baseline C3 performance (AUC: 0.69, 95% CI [0.55–0.82], $p < 0.001$) and baseline PACC5 performance (AUC: 0.75, 95% CI [0.63–0.86], $p = 0.02$).

When looking at the individual C3 measures, 3-month BPSO and FNLT slopes were more strongly related to annual PACC5 change ($r = 0.68$, 95% CI [0.51–0.79], and $r = 0.53$, 95% CI [0.34–0.68] respectively, both $p < 0.001$), compared to the OCL slopes which only reached trend-level significance ($r = 0.21$, 95% CI [−0.02–0.42], $p = 0.07$). Only BPSO 3-month slopes (optimal cut-off: 0.2, AUC: 0.90, 95% CI [0.83–0.97]) showed significantly better discriminative ability than PACC5 baseline scores ($p < 0.001$), whereas the FNLT 3-month slopes (optimal cut-off: 0.4, AUC: 0.80, 95% CI = [0.70–0.90]) and OCL 3-month slopes (optimal cut-off: 0.1, AUC: 0.60, 95% CI = [0.46–0.73]) did not.

DISCUSSION

We demonstrated that CU adults improve over monthly computerized cognitive testing, and that, overall, improvement seems most apparent over initial repeated exposures (i.e., over the first four assessments compared to assessments thereafter). However, individuals vary in their magnitude of improvement over 3 months such that attenuated improvement (i.e., diminished practice effect) was associated with greater global amyloid burden and early tau deposition specifically in the entorhinal cortex. Moreover, 3-month C3 slopes were able to detect differences in spatial tau distribution better than C3 baseline scores. Finally, we showed that the magnitude of C3 slopes over 3 months was predictive of cognitive change over 1 year and could provide a valuable marker to identify individuals who will show more than 0.10 SD annual decline on standard paper-pencil cognitive testing.

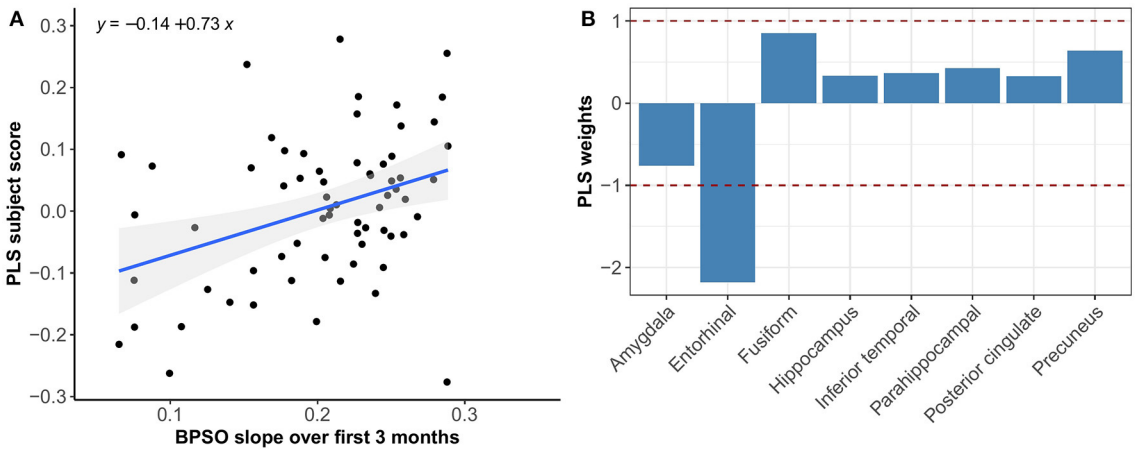


FIGURE 4 | Association between optimal spatial distribution of tau and BPSO slopes over 3 months in the Aβ− group. Left panel **(A)** shows the correlation between PLS subject scores (expression of spatial patterns of tau distribution) and BPSO slopes; right panel **(B)** presents the spatial patterns of tau distribution of z-transformed PLS weights in each ROI.

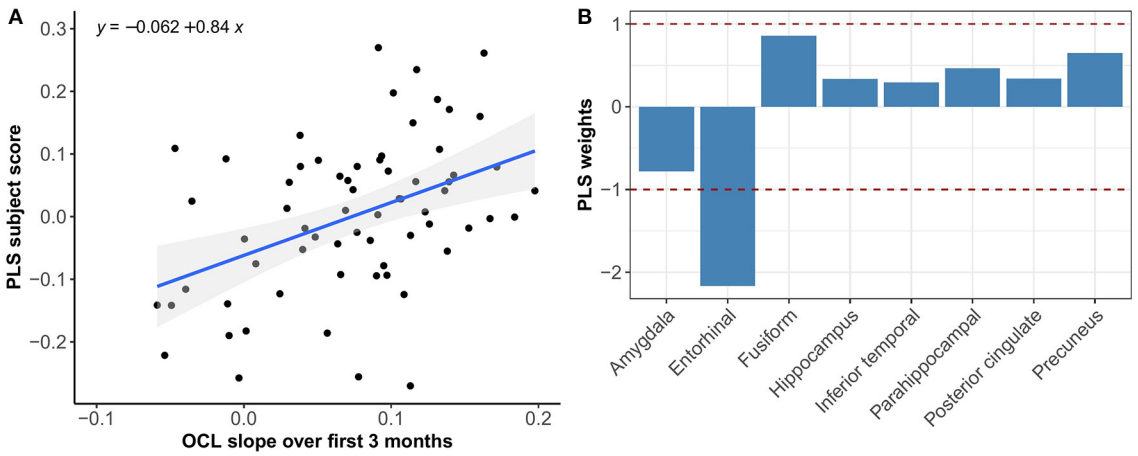


FIGURE 5 | Association between optimal spatial distribution of tau and OCL slopes over 3 months in the Aβ− group. Left panel **(A)** shows the correlation between PLS subject scores (expression of spatial patterns of tau distribution) and OCL slopes; right panel **(B)** presents the spatial patterns of tau distribution of z-transformed PLS weights in each ROI.

Improvement over cognitive testing sessions in the absence of an intervention is thought to reflect practice, also referred to as learning or retest effects, which have typically been viewed as source of error or bias in the context of cognitive testing. There is, however, a growing body of literature suggesting that quantifying PE, and particularly lower or reduced PE, could provide a meaningful clinical marker of (early) subtle decrements in learning and memory performance in preclinical stages of AD (Duff et al., 2007; Hassenstab et al., 2015; Jutten et al., 2020a; Lim et al., 2020; Samaroo et al., 2020). One potential explanation for this is that individuals with reduced PE do not optimally benefit from previous exposure to test material, suggesting worse consolidation and retention of recently learned information induced by deficits in the integrity of their learning and memory system. Impairments in learning and retention of new information have been determined to be the earliest and most robust manifestation of AD, which is in line with consistent observations that AD pathology typically manifests earliest in the MTL and specifically the hippocampal and perirhinal regions that play a crucial role in the learning and consolidation system (Reitz et al., 2009). Our finding that lower PE are associated with greater global amyloid burden and tau deposition in the entorhinal cortex contribute to previous work suggesting that we can potentially capture the first, subtle alterations in learning in preclinical AD by capitalizing on the phenomenon of PE (Samaroo et al., 2020).

PLS results complemented the univariate imaging analyses by showing that 3-month C3 slopes detected differences in spatial tau distribution whereas baseline C3 scores did not. In addition, the PLS analysis revealed that the expression of the

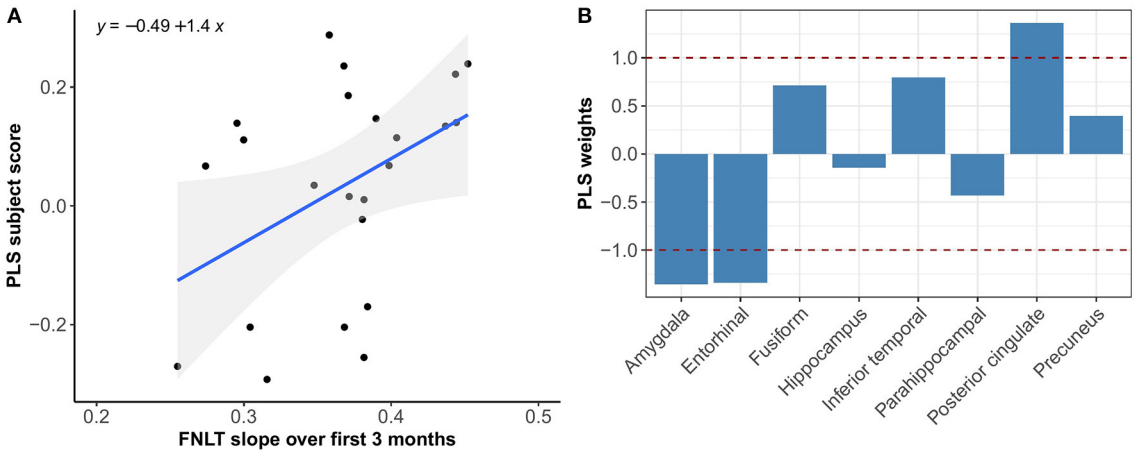


FIGURE 6 | Association between optimal spatial distribution of tau and FNLT slopes over 3 months in the Aβ+ group. Left panel **(A)** shows the correlation between PLS subject scores (expression of spatial patterns of tau distribution) and FNLT slopes; right panel **(B)** presents the spatial patterns of tau distribution of z-transformed PLS weights in each ROI.

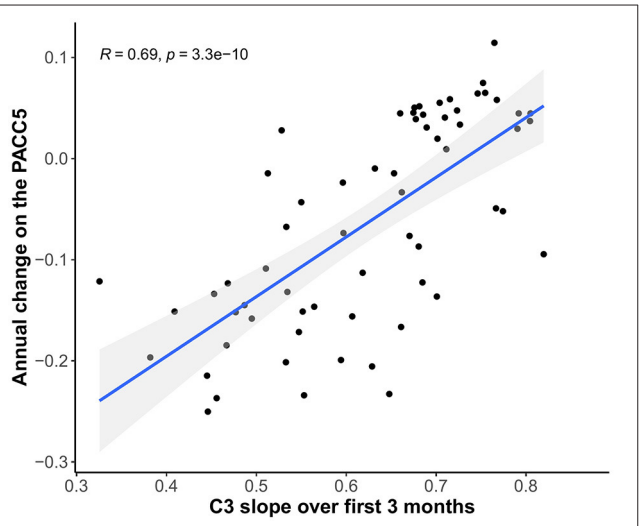


FIGURE 7 | Relationship between C3 slopes over 3 months and annual change scores on the PACC5 (all scores covariate-adjusted).

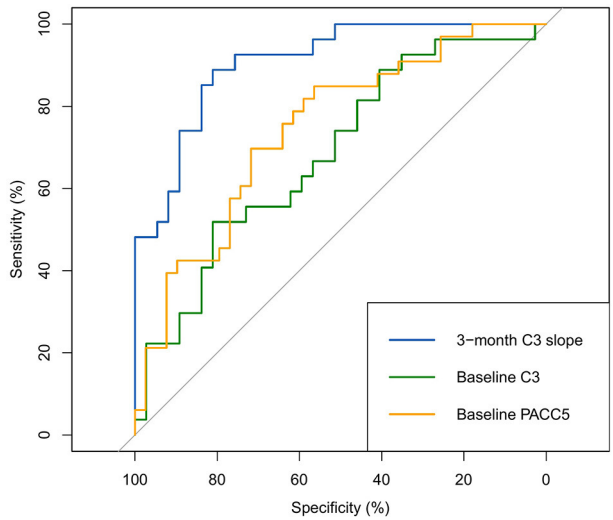


FIGURE 8 | ROC analyses for the identification of >0.10 SD annual decline on the PACC5 (all scores covariate-adjusted).

tau pattern associated with C3 slopes seems different for Aβ− vs. Aβ+ groups. That is, in the Aβ− group we observed a consistent spatial pattern characterized by relatively lower uptake in the entorhinal cortex only, and the expression of this pattern was significantly associated with greater PE over 3 months. In the Aβ+ group we observed a spatial pattern characterized by lower uptake in entorhinal and amygdala and higher uptake in precuneus and posterior cingulate, which was associated with greater PE over 3 months. This difference in spatial patterns could be explained by the fact that Aβ+ group was more progressed in terms of tau pathology than Aβ−, and different relative scales of tau binding in the entorhinal cortex (affected earlier) and posterior regions (affected later) across groups. This is in line with our current understanding that the entorhinal cortex is among the earliest regions of tau accumulation where tau seems to increase with age even before amyloid starts deposing (Maass et al., 2018), whereas other MTL and more posterior regions are affected once amyloid pathology induces the spread of tau into the neocortex (Sanchez et al., 2021). Another interesting difference is that spatial tau distribution in Aβ− was mainly associated with the BPSO and OCL slopes, whereas in the Aβ+ group we only found a significant association with the FNLT slopes. This could be explained by the fact that the nature of the PE observed on the BPSO and OCL (alternate versions) is partly different than the nature of PE on the FNLT (same version), with the latter more heavily relying on remembering the exact

test content which may potentially be affected once amyloid pathology is present.

Besides remembering the exact test content, another potential explanation for the occurrence of PE is that increased familiarity with the test-taking in general leads to the development of strategies and/or reduced test anxiety and stress with repeated testing. These task familiarity effects may be due to procedural learning which is an aspect of cognition that remains relatively spared in early stages of AD (Goldberg et al., 2015). Task familiarity has likely played a role in the PE we observed on particularly the OCL and BPSO, since alternate versions were used for those measures and, thus, ruling out the possibility that PE were only caused by the fact that individuals learned/remembered the specific test items. However, the discrepancy between our findings on the BPSO vs. the OCL regarding the magnitude of PE observed over months and the strength of their associations with annual PACC5 decline, also suggest that test familiarity alone seems insufficient to explain differences in PE. In fact, our findings imply that the nature of the task, as well as the retest paradigm (i.e., same vs. alternate versions) may both contribute to the occurrence and magnitude of PE.

When comparing the different tasks and retest paradigms included in the C3, we found that PE were strongest on the FNLT for which the same version was administered at each time-point. However, PE observed on the BPSO, which was administered using an alternate version retest paradigm (A-B-C-D), showed the strongest sensitivity to differences in early entorhinal tau deposition and better predictive ability for annual decline on the PACC5. The BPSO is a measure of pattern separation, an aspect of episodic memory dependent on hippocampal function (Kirwan and Stark, 2007; Yassa and Stark, 2011) whereby information from overlapping experiences is made independent of one another to overcome interference. Our data showed that task performance relying on this process of pattern separation can improve with practice, even though the individual test items at retest are not the same. This suggests that PE observed on the BPSO are not (only) caused by remembering the exact test items, but that practicing strategies to successfully apply pattern separation plays a role as well. The OCL on the other hand, which was also administered using an alternate retest paradigm and initially designed deliberately to mitigate PE, was less sensitive to PE than both the BPSO and FNLT. This could be explained by the fact that the OCL is a "simpler" measure than the BPSO, providing less room for practicing the required learning/memory strategy, which is in line with previous reports that tests with lower cognitive demand show usually lower PE as opposed to tasks with a larger cognitive demand (Beglinger et al., 2005).

Our finding that PE are most strong with initial repeated exposures is in accordance with a previous meta-analysis and several reviews of PE in the context of longitudinal cognitive aging studies (Beglinger et al., 2005; Calamia et al., 2012; Machulda et al., 2013; Jutten et al., 2020a). These studies show consistently that PE at a group level are most apparent between the first- and second-time testing, and that improvement plateaus after 4–5 repeated assessments. However, it is likely that the moment that individuals reach their plateau differs per individual. Therefore, besides quantifying the amount of improvement over a fixed time-interval as we did in the current study, it would be interesting to characterize learning curves at an individual level and investigate whether the number of assessments needed to reach one's personal plateau could provide an early marker of learning deficits in preclinical AD. Additionally, since other studies have suggested that PE can already be detected over days (Lim et al., 2020) or even over repeated assessments within a single day (Darby et al., 2002), it would be interesting to explore the feasibility and predictive ability of defining even more short-term PE (days rather than months) in the context of preclinical AD (Kaye et al., 2021; Papp et al., 2021a).

Implications

Neuropsychological models have understood the cognitive manifestation of AD in terms of change over years or even decades. However, there is now a developing field that shows how understanding changes in cognition over much shorter periods, such as months, may help inform brain behavior models of the disease, particularly in early or preclinical stages. Our findings provide complementary evidence for the hypothesis that characterizing short-term PE could aid in the detection of individuals at risk for cognitive decline due to AD, above and beyond baseline cognitive scores. This has important implications for clinical trial design and recruitment strategies. First, employing remote, monthly computerized assessments could lead to more rapid recruitment and screening of large samples in a cost-effective manner and maximize sample generalizability by facilitating the inclusion of participants who live in remote locations. Subsequently, characterizing PE over 3 months could advance the more rapid detection of early cognitive change, as well as the identification of those who are at risk for short-term cognitive decline and, thus, may be most likely to benefit from treatment. Ultimately, quantifying PE as a more nuanced way of exploring subtle alterations in cognitive functioning could hopefully increase the rapidity of screening participants and detecting treatment effects in trials that aim slow or halt disease progression in early stages of AD.

Finally, remote cognitive testing may potentially advance the monitoring of (incipient) cognitive impairment in clinical practice. However, not much is known yet about the potential clinical implications of applying a monthly at-home testing paradigm for an individual. For example, the impact of at-home testing on an individual's willingness to have in-clinic follow-up or seek care remains unknown and will thus be an important next step to address in future research.

Study Limitations

An important limitation of the current study is the fact that our study sample is a highly educated and predominantly White cohort, and thereby it is unknown how generalizable our findings are to other populations. Although adherence in our study was high and missing data due to technical difficulties low, it is important to address that a certain level of digital skills as well as compliance to monthly testing are required to successfully

implement these monthly computerized retesting paradigms. The feasibility of at-home computerized testing has previously been demonstrated in HABS and other cohorts (Rentz et al., 2016; Perin et al., 2020), but this should also be determined in more diverse populations and individuals with less digital literacy. Furthermore, it should be noted that we used a study-issued iPad including proprietary CogState software, which may, on the one hand, have contributed to the good adherence, but on the other hand, limits the scalability of the C3.

A general issue with unsupervised cognitive testing is the fact that there is little control over the location, timing of testing, and likelihood of participant distraction, which are all factors likely to interact with task performance and may thereby threaten the internal validity of test scores (Perin et al., 2020). However, a previous study indicated that these factors mainly affect speed of performance rather than accuracy scores (Backx et al., 2020), and since we focused on accuracy measures to detect PE, we do not expect that the uncontrolled environment has biased our findings to a large extent. Furthermore, we applied previously defined cut-offs to ensure the integrity and completion of each individual task, which may also have limited the influence of uncontrolled factors on our results.

A strength of our study is that we complemented univariate imaging analysis with PLS. The main advantage of PLS over univariate analysis is that PLS analysis can examine the relationships between the tau uptake in various regions simultaneously rather than localized tau uptake in each region individually. PLS analysis results are thus expected to be more robust when the input variables are collinear, which is the case with tau uptake in the examined ROI (e.g., uptake in the entorhinal and inferior temporal cortex especially in the Aβ+ group). In addition, PLS analysis may be more robust to noise in the data than univariate analysis. However, it should also be noted that only 8 ROI were included in the PLS analysis. This selection of regions was a-priori defined, based on our initial findings and on what is known about the spread of neocortical tau in cognitively older adults (Johnson et al., 2016; Sanchez et al., 2021). Since our sample consisted mainly of cognitively normal individuals (of which the majority was amyloid-negative), it is expected that there is little or no tau uptake beyond those 8 ROI, and so adding in more regions would likely not benefit our models. An interesting future step would be to investigate the association between PE and tau uptake using voxel-wise analysis. "In addition, it would be worthwhile to use PLS to examine whether regional amyloid accumulation would be associated with the magnitude of PE, especially in amyloid-negative individuals that yet have subthreshold levels of amyloid accumulation (Farrell et al., 2018).

Finally, regarding our investigation of the predictive ability of PE for future cognitive decline, it should be noted that we only had one-year prospective follow-up cognitive testing available. This follow-up duration is particularly short in the context of preclinical AD, which is presumed to be a stage that may last 20 years or longer before the onset of objective cognitive impairment (Sperling et al., 2011). It remains uncertain as to which of our participants would show further cognitive decline and eventually progress to the MCI or dementia stage. Annual data-collection of the HABS cohort is ongoing, which will allow us to address this important question in future research.

CONCLUSION

We showed that, while PE commonly occur in CU adults, diminished PE over monthly computerized cognitive testing are associated with greater AD biomarker burden and cognitive decline over one year. Our findings imply that unsupervised computerized testing using monthly retest paradigms can provide rapid detection of diminished PE indicative of future cognitive decline in preclinical AD. This could aid in more rapid detection of individuals at risk for cognitive decline and thereby accelerate clinical trial recruitment and screening as well as the detection of treatment effects.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Partners Human Research Committee, which is the Institutional Review Board for the Massachusetts General Hospital and Brigham and Women's Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RJ designed and conceptualized the study, analyzed the data, interpreted the data, and drafted the manuscript. JF, DM, and MP assisted with the data analyses and data interpretation. RA, RB, PM, CS, and MY interpreted the data and performed critical editing of the manuscript. DR, KJ, RS, and KP designed and conceptualized the study, interpreted the data, performed critical editing of the manuscript, and provided study supervision. All authors have read and approved the final version of the manuscript.

FUNDING

The Harvard Aging Brain Study was funded by the National Institutes of Health/National Institute on Aging (NIH/NIA) (P01AG036694; Principal Investigators Sperling, Johnson) with additional support from several philanthropic organizations. RJ was supported by a Rubicon grant from the Dutch Research Council (NWO). CS was supported by an R01AG066683-01 award. KP was supported by a K23 award from NIA (1K23AG053422-01).

ACKNOWLEDGMENTS

The authors would like to thank all participants of the Harvard Aging Brain Study, as well as everyone involved in the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2021.800126/full#supplementary-material

REFERENCES

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., and Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the cambridge neuropsychological test automated battery: a within-subjects counterbalanced study. *J. Med. Internet Res.* 22, e16792–e16792. doi: 10.2196/16792

Baker, J. E., Pietrzak, R. H., Laws, S. M., Ames, D., Villemagne, V. L., Rowe, C. C., et al. (2019). Visual paired associate learning deficits associated with elevated beta-amyloid in cognitively normal older adults. *Neuropsychology* 33:964. doi: 10.1037/neu0000561

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch. Clin. Neuropsychol.* 20, 517–529. doi: 10.1016/j.acn.2004.12.003

Buckley, R. F., Sparks, K. P., Papp, K., v, Dekhtyar, M., Martin, C., Burnham, S., et al. (2017). Computerized cognitive testing for use in clinical trials: a comparison of the NIH toolbox and cogstate C3 batteries. *J. Prevent. Alzheimer's Dis.* 4, 3–11. doi: 10.14283/jpad.2017.1

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Dagley, A., LaPoint, M., Huijbers, W., Hedden, T., McLaren, D. G., Chatwal, J. P., et al. (2017). Harvard aging brain study: dataset and accessibility. *NeuroImage* 144, 255–258. doi: 10.1016/j.neuroimage.2015.03.069

Darby, D., Maruff, P., Collie, A., and McStephen, M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. *Neurology* 59, 1042–1046. doi: 10.1212/WNL.59.7.1042

Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., et al. (2014). The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol.* 71, 961–970. doi: 10.1001/jamaneurol.2014.803

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., Haase, R. F., et al. (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch. Clin. Neuropsychol.* 22, 15–24. doi: 10.1016/j.acn.2006.08.013

Duff, K., Callister, C., Dennett, K., and Tometich, D. (2012). Practice effects: a unique cognitive variable. *Clin. Neuropsychol.* 26, 1117–1127. doi: 10.1080/13854046.2012.722685

Farrell, M. E., Chen, X., Rundle, M. M., Chan, M. Y., Wig, G. S., and Park, D. C. (2018). Regional amyloid accumulation and cognitive decline in initially amyloid-negative adults. *Neurology* 91, e1809–e1821. doi: 10.1212/WNL.0000000000006469

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1037/t07757-000

Gold, M., Amatniek, J., Carrillo, M. C., Cedarbaum, J. M., Hendrix, J. A., Miller, B. B., et al. (2018). Digital technologies as biomarkers, clinical outcomes assessment, and recruitment tools in Alzheimer's disease clinical trials. *Alzheimer's Dement. (N Y).* 4, 234–242. doi: 10.1016/j.trci.2018.04.003

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

Grober, E., Ocepek-Welikson, K., and Teresi, J. A. (2009). The free and cued selective reminding test: evidence of psychometric adequacy. *Psychol. Sci. Q.* 51, 266–282.

Hanseeuw, B. J., Betensky, R. A., Jacobs, H. I., Schultz, A. P., Sepulcre, J., Becker, J. A., et al. (2019). Association of amyloid and tau with cognition in preclinical Alzheimer's disease: a longitudinal study. *JAMA Neurol.* 76, 915–924. doi: 10.1001/jamaneurol.2019.1424

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Johnson, K. A., Schultz, A., Betensky, R. A., Becker, J. A., Sepulcre, J., Rentz, D., et al. (2016). Tau positron emission tomographic imaging in aging and early Alzheimer's disease. *Ann. Neurol.* 79, 110–119. doi: 10.1002/ana.24546

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S. E., et al. (2020a). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimer's Dement.* 12:e12055. doi: 10.1002/dad2.12055

Jutten, R. J., Sikkes, S. A. M., Amariglio, R. E., Buckley, R. F., Properzi, M. J., Marshall, G. A., et al. (2020b). Identifying sensitive measures of cognitive decline at different clinical stages of Alzheimer's disease. *J. Int. Neuropsychol. Soc.* 27, 426–438. doi: 10.1017/S1355617720000934

Kaye, J., Aisen, P., Amariglio, R., Au, R., Ballard, C., Carrillo, M., et al. (2021). Using digital tools to advance Alzheimer's drug trials during a pandemic: the EU/US CTAD task force. *J. Prevent. Alzheimer's Dis.* 8, 513–519. doi: 10.14283/jpad.2021.36

Kirwan, C. B., and Stark, C. E. L. (2007). Overcoming interference: an fMRI investigation of pattern separation in the medial temporal lobe. *Learn. Mem.* 14, 625–633. doi: 10.1101/lm.663507

Koo, B. M., and Vizer, L. M. (2019). Mobile technology for cognitive assessment of older adults: a scoping review. *Innov. Aging* 3, 1–14. doi: 10.1093/geroni/igy038

Lee, C. M., Jacobs, H. I. L., Marqui,é, M., Becker, J. A., Andrea, N., v, Jin, D. S., et al. (2018). 18F-flortaucipir binding in choroid plexus: related to race and hippocampus signal. *J. Alzheimer's Dis.* 62, 1691–1702. doi: 10.3233/JAD-170840

Lim, Y. Y., Baker, J. E., Bruns, L., Mills, A., Fowler, C., Fripp, J., et al. (2020). Association of deficits in short-term learning and Aβ and hippocampal volume in cognitively normal adults. *Neurology* 95, e2577–e2585. doi: 10.1212/WNL.0000000000010728

Lim, Y. Y., Ellis, K. A., Harrington, K., Ames, D., Martins, R. N., Masters, C. L., et al. (2012). Use of the CogState Brief Battery in the assessment of Alzheimer's disease related cognitive impairment in the Australian Imaging, Biomarkers and Lifestyle (AIBL) study. *J. Clin. Exp. Neuropsychol.* 34, 345–358. doi: 10.1080/13803395.2011.643227

Maass, A., Lockhart, S. N., Harrison, T. M., Bell, R. K., Mellinger, T., Swinnerton, K., et al. (2018). Entorhinal tau pathology, episodic memory decline, and neurodegeneration in aging. *J. Neurosci.* 38, 530–543. doi: 10.1523/JNEUROSCI.2028-17.2017

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

Maruff, P., Thomas, E., Cysique, L., Brew, B., Collie, A., Snyder, P., et al. (2009). Validity of the CogState brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Arch. Clin. Neuropsychol.* 24, 165–178. doi: 10.1093/arclin/acp010

McIntosh, A. R., Bookstein, F. L., Haxby, J., v, and Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage* 3, 143–157. doi: 10.1006/nimg.1996.0016

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Arch. Neurol.* 49, 1253–1258. doi: 10.1001/archneur.1992.00530360051017

Mormino, E. C., Betensky, R. A., Hedden, T., Schultz, A. P., Ward, A., Huijbers, W., et al. (2014). Amyloid and emAPOE ε4/em interact to influence short-term decline in preclinical Alzheimer's disease. *Neurology* 82, 1760 LP—1767. doi: 10.1212/WNL.0000000000000431

Mortamais, M., Ash, J. A., Harrison, J., Kaye, J., Kramer, J., Randolph, C., et al. (2017). Detecting cognitive changes in preclinical Alzheimer's disease: a review of its feasibility. *Alzheimer's Dement.* 13, 468–492. doi: 10.1016/j.jalz.2016.06.2365

Papp, K., v, Rentz, D. M., Orlovsky, I., Sperling, R. A., and Mormino, E. C. (2017). Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: the PACC5. *Alzheimer's Dement.* 3, 668–677. doi: 10.1016/j.trci.2017.10.004

Papp, K. V., Buckley, R., Mormino, E., Maruff, P., Villemagne, V. L., Masters, C. L., et al. (2020). Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. *Alzheimer's Dement.* 16, 552–560. doi: 10.1016/j.jalz.2019.09.074

Papp, K. V., Rentz, D. M., Buckley, R. F., Schneider, O. R., Hsieh, S., Soberanes, D., et al. (2021a). "Associations between remote cognitive testing on an individual's own digital device and amyloid burden on neuroimaging in clinically normal older adults: results from Boston Remote Assessment for Neurocognitive Health (BRANCH)," in *2021 Alzheimer's Association International Conference (ALZ)* (San Diego, CA).

Papp, K. V., Rentz, D. M., Maruff, P., Sun, C.-K., Raman, R., Donohue, M. C., et al. (2021b). The computerized cognitive composite (c3) in a4, an Alzheimer's disease secondary prevention trial. *J. Prevent. Alzheimer's Dis.* 8, 59–67. doi: 10.14283/jpad.2020.38

Perin, S., Buckley, R. F., Pase, M. P., Yassi, N., Lavale, A., Wilson, P. H., et al. (2020). Unsupervised assessment of cognition in the Healthy Brain Project: implications for web-based registries of individuals at risk for Alzheimer's disease. *Alzheimer's Dement.* 6:e12043. doi: 10.1002/trc2.12043

Petersen, R. C., Wiste, H. J., Weigand, S. D., Fields, J. A., Geda, Y. E., Graff-Radford, J., et al. (2021). NIA-AA Alzheimer's disease framework: clinical characterization of stages. *Ann. Neurol.* 89, 1145–1156. doi: 10.1002/ana.26071

Petersen, R. C., Wiste, H. J., Weigand, S. D., Rocca, W. A., Roberts, R. O., Mielke, M. M., et al. (2016). Association of elevated amyloid levels with cognition and biomarkers in cognitively normal people from the community. *JAMA Neurol.* 73, 85–92. doi: 10.1001/jamaneurol.2015.3098

Reitz, C., Honig, L., Vonsattel, J. P., Tang, M. X., and Mayeux, R. (2009). Memory performance is related to amyloid and tau pathology in the hippocampus. *J. Neurol. Neurosurg. Psychiatry* 80, 715–721. doi: 10.1136/jnnp.2008.154146

Rentz, D. M., Amariglio, R. E., Becker, J. A., Frey, M., Olson, L. E., Frishe, K., et al. (2011). Face-name associative memory performance is related to amyloid burden in normal elderly. *Neuropsychologia* 49, 2776–2783. doi: 10.1016/j.neuropsychologia.2011.06.006

Rentz, D. M., Dekhtyar, M., Sherman, J., Burnham, S., Blacker, D., Aghjayan, S. L., et al. (2016). The feasibility of At-Home iPad cognitive testing for use in clinical trials. *J. Prevent. Alzheimer's Dis.* 3, 8–12. doi: 10.14283/jpad.2015.78

Salthouse, T. A. (2012). Robust cognitive change. *J. Int. Neuropsychol. Soc.* 18, 749–756. doi: 10.1017/S1355617712000380

Samaroo, A., Amariglio, R. E., Burnham, S., Sparks, P., Properzi, M., Schultz, A. P., et al. (2020). Diminished learning over repeated exposures (LORE) in preclinical Alzheimer's disease. *Alzheimer's Dement.* 12:e12132. doi: 10.1002/dad2.12132

Sanchez, J. S., Becker, J. A., Jacobs, H. I., Hanseeuw, B. J., Jiang, S., Schultz, A. P., et al. (2021). The cortical origin and initial spread of medial temporal tauopathy in Alzheimer's disease assessed with positron emission tomography. *Sci. Transl. Med.* 13:eabc0655. doi: 10.1126/scitranslmed.abc0655

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Stark, S. M., Yassa, M. A., Lacy, J. W., and Stark, C. E. L. (2013). A task to assess behavioral pattern separation (BPS) in humans: data from healthy aging and mild cognitive impairment. *Neuropsychologia* 51, 2442–2449. doi: 10.1016/j.neuropsychologia.2012.12.014

Stricker, N. H., Lundt, E. S., Alden, E. C., Albertson, S. M., Machulda, M. M., Kremers, W. K., et al. (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. *J. Prevent. Alzheimer's Dis.* 7, 21–28. doi: 10.14283/jpad.2019.35

Villemagne, V. L., Pike, K. E., Chételat, G., Ellis, K. A., Mulligan, R. S., Bourgeat, P., et al. (2011). Longitudinal assessment of Aβ and cognition in aging and Alzheimer's disease. *Ann. Neurol.* 69, 181–192. doi: 10.1002/ana.22248

Watson, F. L., Pasteur, M. L., Healy, D. T., and Hughes, E. A. (1994). Nine parallel versions of four memory tests: an assessment of form equivalence and the effects of practice on performance. *Hum. Psychopharmacol. Clin. Exp.* 9, 51–61. doi: 10.1002/hup.470090107

Wechsler, D. (1987). *WMS-R: Wechsler Memory Scale-Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Yassa, M. A., and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends Neurosci.* 34, 515–525. doi: 10.1016/j.tins.2011.06.006

Conflict of Interest: RS receives research funding from NIA, Alzheimer's Association, Eli Lilly and Co., and Eisai. She has served as a consultant to AC Immune, Alyn Cytox, Janssen, Neurocentria, Roche, Prothena, and Shionogi but not directly relevant to this study. MY receives research funding from NIA and NIMH, served as a consultant for Pfizer, Eisai, Cognito Therapeutics, Curasen Therapeutics, BPT Pharma and Dart Neuroscience, and is co-founder of Enthorin Therapeutics and Augnition Labs, none of which are directly relevant to this study. PM is employed full time by Cogstate Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 13 April 2022
doi: 10.3389/fnagi.2022.863942

Episodic Memory and Executive Function Are Differentially Affected by Retests but Similarly Affected by Age in a Longitudinal Study of Normally-Aging Older Adults

Elizabeth L. Glisky[1], Cindy B. Woolverton[1], Katelyn S. McVeigh[2] and Matthew D. Grilli[2]**

[1] Aging and Cognition Laboratory, Department of Psychology, University of Arizona, Tucson, AZ, United States, [2] Human Memory Laboratory, Department of Psychology, University of Arizona, Tucson, AZ, United States

Episodic memory and executive function are two cognitive domains that have been studied extensively in older adults and have been shown to decline in normally-aging older individuals. However, one of the problems with characterizing cognitive changes in longitudinal studies has been separating effects attributable to normal aging from effects created by repeated testing or practice. In the present study, 166 people aged 65 and older were enrolled over several years and tested at least 3 times at variable intervals (M = 3.2 yrs). The cognitive measures were composite scores. Each composite was made up of five neuropsychological tests, previously identified through factor analysis. For one pair of composite scores, variance attributable to age was removed from each subtest through regression analyses before z-scores were computed, creating two age-corrected composites. A second pair of composites were not age-corrected. Using linear mixed-effects models, we first explored retest effects for each cognitive domain, independent of age, using the age-corrected composites. We then modeled aging effects using the age-uncorrected composites after subtracting out retest effects. Results indicated significant retest effects for memory but not for executive function, such that memory performance improved across the three testing sessions. When these practice effects were removed from the age-uncorrected data, effects of aging were evident for both executive and memory function with significant declines over time. We also explored several individual difference variables including sex, IQ, and age at the initial testing session and across time. Although sex and IQ affected performance on both cognitive factors at the initial test, neither was related to practice effects, although young-older adults tended to benefit from practice to a greater extent than old-older adults. In addition, people with higher IQs showed slower age-related declines in memory, but no advantages in executive function. These findings suggest that (a) aging affects both memory and executive function similarly, (b) higher IQ, possibly reflecting cognitive reserve, may slow age-related declines in memory, and (c) practice through repeated testing enhances performance in memory particularly in younger-older adults, and may therefore mask aging effects if not taken into account.

Keywords: aging, episodic memory, executive function, longitudinal, retests, practice effects

OPEN ACCESS

Edited by:
Lars Nyberg,
Umeå University, Sweden

Reviewed by:
Eero Vuoksimaa,
University of Helsinki, Finland
Elena Rodriguez-Vieitez,
Karolinska Institutet (KI), Sweden

****Correspondence:***
Elizabeth L. Glisky
glisky@arizona.edu
Matthew D. Grilli
mdgrilli@arizona.edu

Specialty section:
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

Received: *27 January 2022*
Accepted: *22 March 2022*
Published: *13 April 2022*

Citation:
Glisky EL, Woolverton CB, McVeigh KS and Grilli MD (2022) Episodic Memory and Executive Function Are Differentially Affected by Retests but Similarly Affected by Age in a Longitudinal Study of Normally-Aging Older Adults. Front. Aging Neurosci. 14:863942. doi: 10.3389/fnagi.2022.863942

INTRODUCTION

Longitudinal studies of cognitive function in older adults have a relatively recent history, with the bulk of the research appearing in the literature since 2000. Much of this work initially focused on memory and speed of processing, areas of cognition that showed clear age-related differences in cross-sectional studies. More recent longitudinal studies have included other cognitive domains including executive function (e.g., Gross et al., 2015; Hassenstab et al., 2015), but in most studies, little has been said concerning what specific cognitive processes within a domain might be implicated in changes over time. In addition, studies have begun relating age-related cognitive changes to corresponding brain changes (e.g., Kramer et al., 2007; Persson et al., 2014; Armstrong et al., 2020; Gavett et al., 2021). Results of these longitudinal studies, however, have not been entirely consistent with respect to the cognitive domains most affected by age, the rate of decline over time, and the variables that might moderate change. Several researchers (e.g., Rabbitt et al., 2001; Ferrer et al., 2004, 2005; Rönnlund et al., 2005; Wilson et al., 2006; Salthouse, 2010) have also acknowledged that repeated testing can influence and thereby mask age-related changes, and have proposed different ways of accounting for and eliminating such effects. Although practice effects are usually greatest after short intervals, some studies have reported effects even 5-6 years following initial testing (e.g., Elman et al., 2018). It has also been suggested that practice effects themselves might reveal important individual differences in the cognitive functioning of older people (e.g., Machulda et al., 2013; Hassenstab et al., 2015).

For the present longitudinal study, we looked at composite measures of episodic memory and executive function in a sample of normally-aging older adults. Tests comprising each cognitive domain were chosen to reflect a common process, and the makeup of the composites was derived through factor analyses. We incorporated a novel way to separate aging and practice effects, and explored the impact of several individual difference variables on both retesting and aging.

We first began gathering neuropsychological test data from older adults in 1992 in the context of studying source memory. At that time, some studies had shown that on occasions when amnesic patients with medial temporal lobe damage recalled a recently-presented fact, they could also recall its source —where they heard it or who told them (e.g., Schacter et al., 1984). On the other hand, patients with damage to the frontal lobes, who could readily recall the facts, often could not recall their source (e.g., Janowsky et al., 1989). The two kinds of memory thus appeared to depend on different brain regions—recently presented fact memory on medial temporal lobe structures, and source memory on frontal brain structures. Subsequent studies reporting source memory deficits in older adults, further suggested that these deficits might indicate declining frontal lobe function in older people (e.g., Craik et al., 1990), but findings were inconsistent.

To test this hypothesis in older adults (Glisky et al., 1995), we chose tests from our neuropsychological battery thought to depend on each brain region. Specifically, we selected tests of episodic memory that varied in stimulus properties (e.g., verbal, visual, facial), encoding processes (e.g., single items, pairs, narratives), and retrieval processes (i.e., free recall, cued recall, recognition), but shared processes involved in the fundamental retention or consolidation of information over time, processes dependent on the medial temporal lobes. On the other hand, tests of executive function, thought to depend primarily on the frontal lobes, were selected to reflect control processes that were not involved in episodic memory, but instead were thought to be similar to executive processes associated with working memory. This assumption was supported in a later study by McCabe et al. (2010), who reported a high correlation (0.96) between our executive function composite (minus one common test) and a composite measure made up of complex span tasks.

To verify that these tests were indeed measuring separate constructs, we conducted a series of factor analyses. Because we were interested in the differential contributions of neurocognitive processes that were independent of age, variance attributable to age was removed from each individual test through regression analyses, and the residual scores were then submitted to factor analysis. The initial principal components analysis revealed two independent and uncorrelated factors. Composite factor scores, representing the average of the component test *z*-scores (equally weighted), were then assigned to each individual. Two later confirmatory factor analyses on separate and larger groups of older adults confirmed the two-factor solution and several follow-up studies showed that the two factors were differentially associated with item and source memory in older adults (Glisky et al., 2001; Glisky and Kong, 2008).

Rather than re-calculate and re-assign z-scores for each study sample going forward, we created a standardized reference group based on 227 community-dwelling older adults, who received these same tests, between 1998 and 2004. The data from this group then provided the means, standard deviations, and age corrections for classifying all past and future participants with respect to their episodic memory and executive function. We also created a parallel set of scores without the age correction, for studies in which age was a variable of interest (e.g., Glisky and Kong, 2008). The reference group, aged 65-90 (M = 73.4), had a mean education level of 15.6 years, were in good health, were not depressed or taking anti-depressant medications, reported no previous psychiatric or neurological problems that might have affected cognitive function, and had a score $\geq$ **26** (**M** = 28.9) on the Mini-Mental State Examination (MMSE; Folstein et al., 1975). As our experimental studies continued over time, several people who had participated in our previous studies returned, and were re-tested to ensure that their cognitive profiles were up-to-date. Although not our primary goal at the time, this enabled the collection of longitudinal data, which, after several years, has allowed us to look at longitudinal changes in episodic memory and executive function and to contribute to this special issue on the importance of cognitive practice effects in aging neuroscience.

There are many reasons why practice effects should be considered in longitudinal studies of aging, many or all of which we expect will be addressed in this special issue. Our interests lay specifically in documenting and understanding the processes involved in "normal" cognitive aging, but because of repeated testing of the same materials and/or procedures, this was not a straightforward matter. Practice effects could elevate

performance, making it difficult to assess the actual extent of normal aging processes. Several questions about practice could, and have been asked, including (a) Do all cognitive functions benefit equally from practice, and (b) what individual difference factors might influence practice effects? These were questions that we hoped to address with our data. In addition, understanding variability in the effects of practice might not only provide a greater understanding of the normal aging processes and the cognitive functions most affected, it might also help us in future studies to identify those individuals who were not aging normally, and perhaps suggest intervention strategies.

On the basis of prior studies, we expected that our episodic memory factor would show improved performance across testing sessions, and declines with increasing age. The few studies that have included measures of executive function and working memory have been inconclusive with respect to both retest and aging effects, and so suggested no clear hypotheses with respect to the executive function factor.

MATERIALS AND METHODS

Participants

The present study includes data for 166 older adults between the ages of 65 and 91, who completed at least three testing sessions, were recruited continuously over a period of 18 years, and were retested at varying intervals (M = 3.2 yrs, SD = 1.4). The recruitment of participants for initial testing was conducted through the distribution of fliers in the local community, advertisements in the local paper, and public talks to groups at senior centers. Although some individuals continued to return for further tests (e.g., 83 people had at least 4 tests), we will focus here on the first three test sessions for which we have complete data. To ensure that our sample continued to warrant the label "normally-aging older adults," we retained the exclusion criteria that we used for our standard reference group (see above), and removed people from the longitudinal study if they failed to meet those criteria in any of their test sessions. Those whose composite scores for either of the cognitive domains fell to more than 2 SDs below the mean were also dropped from further participation. Of the 547 older people who completed initial neuropsychological testing between 1992 and 2010, 53% (N = 292) completed Time 2 testing, and 58% of those individuals completed the third session. People failed to continue for a variety of reasons. Most dropouts were attributable to lost contact, lost interest, or ongoing physical or medical limitations. Of the 255 people who dropped between Test 1 and Test 2, 99 failed to meet inclusion criteria; 14 of those had neuropsychological scores more than 2 SDs below the mean, and 3 had MMSE scores below 26. Fifteen people failed to meet inclusion criteria for Test 3, two of whom had low MMSE scores. Three people were subsequently excluded because of missing FSIQ scores. Overall, those who dropped out tended to be older and had lower cognitive scores. All older adults in the present study, 114 women and 52 men, continued to perform within normal limits throughout all test sessions. Their mean age at Test 1 was 71.7 years (SD = 4.8), mean education 16.0 years (SD = 2.5), and mean MMSE score 29.1 (SD = 1.0). All studies that contributed data to the present study and their corresponding consent forms were approved by the University of Arizona's Human Subjects Protection Program. Written informed consent was obtained on each testing occasion.

Cognitive Tests and Measures

The primary outcome measures were the composite z-scores representing performance on the two uncorrelated neurocognitive factors, each derived from five neuropsychological tests. Tests contributing to the executive function (EF) factor included the number of categories achieved on the Modified Wisconsin Card Sorting Test (Hart et al., 1988), the total number of words produced to the cues F, A, and S in a verbal fluency task (Spreen and Benton, 1977), Backward Digit Span and Mental Control from the Wechsler Memory Scale-R or III (Wechsler, 1987, 1997b), and Mental Arithmetic from the Wechsler Adult Intelligence Scale-R (Wechsler, 1981). Tests representing episodic memory function (MF) included Logical Memory I, Verbal Paired Associates 1 and Faces 1 all from Wechsler Memory Scale-R or III (Wechsler, 1987, 1997b), Visual Paired Associates II from Wechsler Memory Scale-R (Wechsler, 1987), and Long-Delay Cued Recall from the California Verbal Learning Test (Delis et al., 1987). Two z-scores were assigned to each participant for each cognitive factor, one representing age-corrected performance and the other age-uncorrected performance. Participants also completed IQ tests, the full tests prior to 1999 (Wechsler, 1981, 1997a) and the abbreviated version thereafter (Wechsler, 1999). **Table 1** shows that at baseline (Test 1), individuals in the present study were on average 1.7 years younger than the reference group and performed at a somewhat higher level on the cognitive tests.

Data Analysis

Practice Effects

We used linear mixed effects models to examine the longitudinal relation between repeated testing (1, 2, and 3) and age-corrected EF and MF scores. As noted above, variance attributable to age had been removed from these scores, eliminating any effects of increasing age across tests. The models included test session (centered such that test session 1 was the intercept) as our predictor of practice effects. The coefficient for test session reflects the longitudinal effect of repeated testing for each additional test session. To examine individual differences in the rate of change associated with one more test session, we also included age at baseline, sex, and baseline FSIQ, and their interactions with test session. We centered baseline age at 72 years, which was the round number closest to the average baseline age of the cohort. FSIQ was centered at the round number closest to the average FSIQ at baseline for the sample, which was 124. We included random intercepts in these models. Because test sessions 2 and 3 did not occur at fixed time intervals, we also ran the models examining practice effects on age-corrected EF and MF scores including two additional predictors: years since baseline, and the interaction between years since baseline and test session. However, these predictors were not significant in either model, and model comparison indicated that including them did not significantly improve model fit. For

parsimony, we therefore did not include them in the final models examining practice effects.

Aging Effects

We applied the same linear mixed effects modeling approach to evaluate the role of increasing age on practice-corrected EF and MF scores. In these models, the primary predictor was years since baseline or "time," to capture the effects of aging. The coefficient for time reflects the longitudinal effect of one more year of age on the cognitive outcomes. Whereas age-corrected scores were used to examine practice effects, practice-corrected scores were used to examine age effects. To derive these practice-corrected scores, we calculated the absolute difference between the age-corrected z-scores at session 1 and 2, and session 2 and 3, and subtracted the relevant difference scores from the age-uncorrected *z*-scores at session 2 or 3. Conceptually, this approach assumes that the differences between the age-corrected z-scores in session 1 and 2, and 2 and 3, primarily reflects the effects of practice, which are then removed from the age-uncorrected z-scores, creating the practice-free *z*-scores. These models also included age at baseline, sex, and baseline FSIQ, and their interactions with time to determine whether they influenced the age-related decline. As before, random intercepts were included.

RStudio was used for statistical analyses and data visualization (R Core Team, 2019), including lme4 (Bates et al., 2015), lmertest (Kuznetsova et al., 2017) to calculate p values, and "ggplot2" for data visualization (Wickham, 2016).

RESULTS

Practice Effects

Mean *z*-scores for the two cognitive factors across the three test sessions are shown in **Table 1**. These composite measures are age-corrected such that increases in age over time cannot affect any boost in scores attributable to retesting. The data indicate little change in EF scores with repeated testing, but a substantial increase in performance on the MF tests. Individual data are shown in **Figure 1**. In this figure, each individual's performance is represented by a thin blue (EF Factor) or purple (MF Factor) line, and the longitudinal effects of repeated testing from the linear mixed effects models described below are overlaid on these raw cognitive composite scores. Note that the interval between test sessions is variable across individuals (M = 3.2 yrs), and so does not represent continuous time.

For EF (**Figure 1A**), age-corrected z-scores actually showed small but non-significant decreases with each repeated test (β = −0.027, SE = 0.019, p = 0.144), suggesting an absence of practice effects. This non-significant decline in EF scores was moderated by baseline age, as indicated by a significant interaction between baseline age and test session (β = −0.011, SE = 0.003, $p < 0.001$), but was not affected by FSIQ (β = 0.001, SE = 0.001, p = 0.306) or sex (β = −0.011, SE = 0.033, p = 0.729). As shown in **Figure 2**, although individuals older than the mean of 72 years on average (i.e., the old-older group (+1 SD = 5 yrs) showed a significant decline over test session (β = −0.079, SE = 0.025, p = 0.001), individuals at the mean or younger (i.e., the young-older group) on average showed neither a significant increase or decrease across test sessions (mean age: β = −0.024, SE = 0.019, p = 0.193; 1 SD below mean age: β = 0.031, SE = 0.024, p = 0.194). There was therefore no evidence of significant practice effects in EF in any age group.

On the other hand, for MF (**Figure 1B**), age-corrected z-scores showed clear and significant increases with each repeated test (β = 0.148, SE = 0.021, $p < 0.001$), reflecting practice effects. Here too, practice effects were significantly moderated by baseline age as reflected in the significant interaction with test session (β = −0.011, SE = 0.004, p = 0.002), but not by FSIQ (β = 0.002, SE = 0.001, p = 0.121) or sex (β = −0.030, SE = 0.038, p = 0.426). However, as shown in **Figure 3**, there were robust benefits of practice for MF scores regardless of baseline age (1 SD older than the mean: β = 0.097, SE = 0.028, $p < 0.001$; mean age: β = 0.151, SE = 0.021, $p < 0.001$; 1 SD younger than the mean: β = 0.204, SE = 0.027, $p < 0.001$). These practice effects, however, were smaller in those who were older on average at baseline, accounting for the interaction. Note also that preliminary analyses found no effect of time since baseline on practice effects, indicating that at long intervals (> 2 yrs), the number of years since the prior test did not predict practice effects.

Although only baseline age affected practice across testing sessions in either cognitive function, all of the individual difference variables contributed to cross-sectional differences in performance at baseline. For EF, there was a significant effect of baseline age (β = 0.018, SE = 0.008, p = 0.029), indicating that being older than 72 at baseline was associated with higher EF scores at the initial testing session. There was also a significant effect of FSIQ (β = 0.025, SE = 0.003, $p < 0.001$): Individuals with higher intelligence had higher baseline EF scores. Finally, there was a significant effect of sex, such that men had higher baseline EF scores than women (β = 0.203, SE = 0.087, p = 0.021).

TABLE 1 | Mean (sd) age-corrected composite z-scores, age, and FSIQ for reference group and study sample.

	Reference Group N = 227	Study Sample N = 166		
		Test 1	Test 2	Test 3
Age-corrected EF Factor	−0.0006 (0.66)	0.13 (0.62)	0.15 (0.62)	0.07 (0.66)
Age-Corrected MF Factor	−0.006 (0.63)	0.18 (0.62)	0.36 (0.63)	0.46 (0.60)
Age (yrs)	73.4 (5.4)	71.7 (4.8)	75.0 (4.9)	78.1 (5.0)
FSIQ	122.7 (13.8)	124.1 (12.2)	124.3 (11.4)	125.3 (12.1)

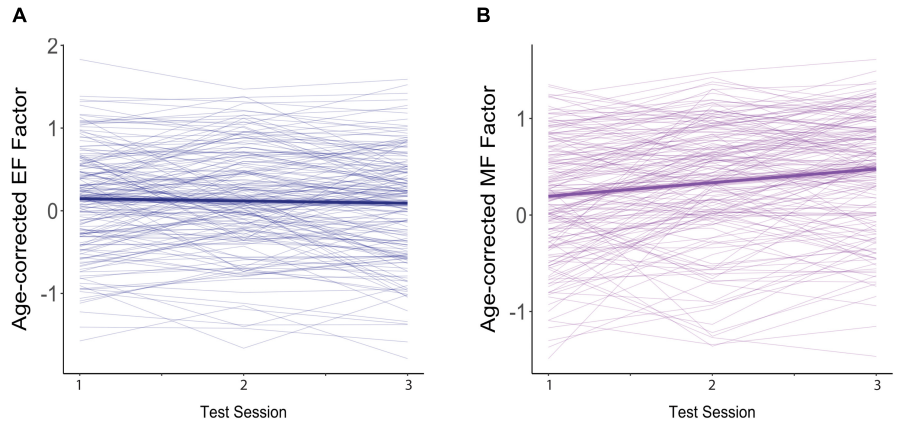


FIGURE 1 | Effects of practice on age-corrected factor scores for **(A)** executive function (EF) and **(B)** memory function (MF). The dark blue (EF) and purple (MF) lines reflect the overall trend across test sessions. The colored ribbon around these lines is the 95% confidence interval. Each participant's scores across the three test sessions are connected by a thin, light-colored line.

For MF, there was also a significant cross-sectional effect of baseline age (β = 0.015, *SE* = 0.008, $p = 0.044$), indicating that being older than 72 at baseline was associated with higher baseline MF scores. There again was a significant effect of FSIQ (β = 0.020, *SE* = 0.003, $p < 0.001$): Individuals with higher intelligence had higher baseline MF scores. Finally, there was a significant effect of sex, such that women had higher baseline MF scores than men (β = −0.606, *SE* = 0.080, $p < 0.001$).

Aging Effects

The effects of aging on our two cognitive factors are shown in **Figure 4**. For these analyses, we used practice-corrected EF and MF scores (regardless of whether there were significant effects of repeated testing) to ensure that age effects were not masked by practice effects. The longitudinal effects of time from the models below are overlaid on these raw cognitive composite scores.

For both EF (**Figure 4A**) and MF (**Figure 4B**), practice-corrected z-scores significantly decreased as time passed, indicating age-related cognitive decline in both cognitive domains (EF: β = −0.071, *SE* = 0.007, $p < 0.001$; MF: β = −0.041, *SE* = 0.007, $p < 0.001$).

For EF scores, decline over time was significantly moderated by baseline age (*B* = −0.005, *SE* = 0.001, $p < 0.001$), but not by FSIQ (β = −0.0009, *SE* = 0.0005, $p = 0.104$) or sex (β = −0.00007, *SE* = 0.014, $p = 0.996$). As shown in **Figure 5**, practice-corrected EF scores significantly decreased over time regardless of baseline age (1 *SD* older than the mean: β = −0.094, *SE* = 0.010, $p < 0.001$; mean age: β = −0.07, *SE* = 0.007, $p < 0.001$; 1 *SD* younger than the mean: β = −0.046, *SE* = 0.010, $p < 0.001$), but the rate of decline

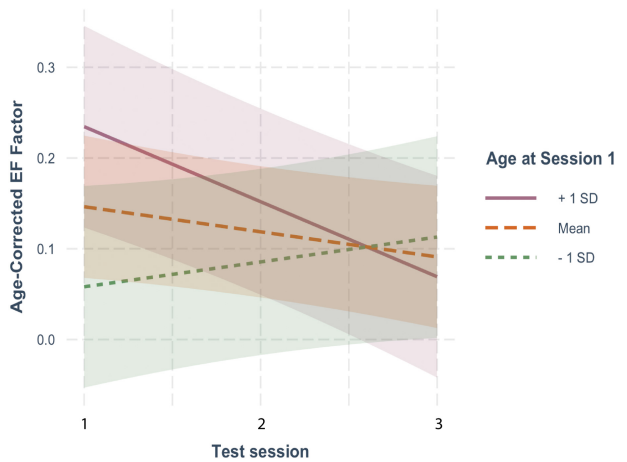


FIGURE 2 | The moderating effect of baseline age on the rate of change in age-corrected EF scores across sessions. The solid line represents individuals who were on average 1 SD older than the mean age at baseline; the large-dashed line shows the performance of the mean age group, and the small-dashed line portrays those 1 SD younger than the mean age The colored ribbon around each line is the 95% confidence interval.

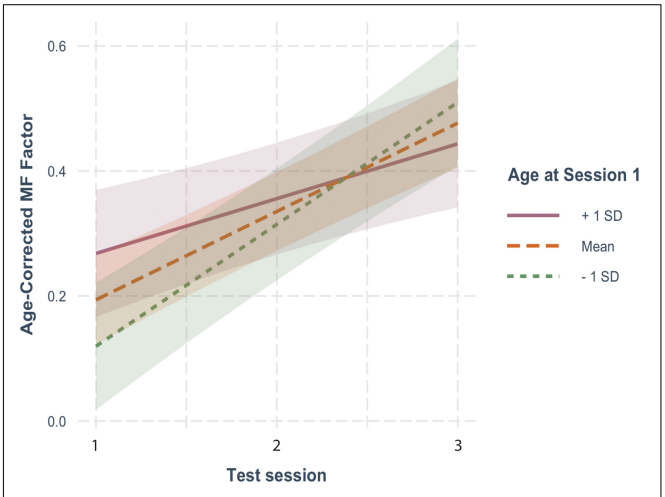


FIGURE 3 | The moderating effect of baseline age on the rate of change in age-corrected MF scores across sessions. See **Figure 2** for details.

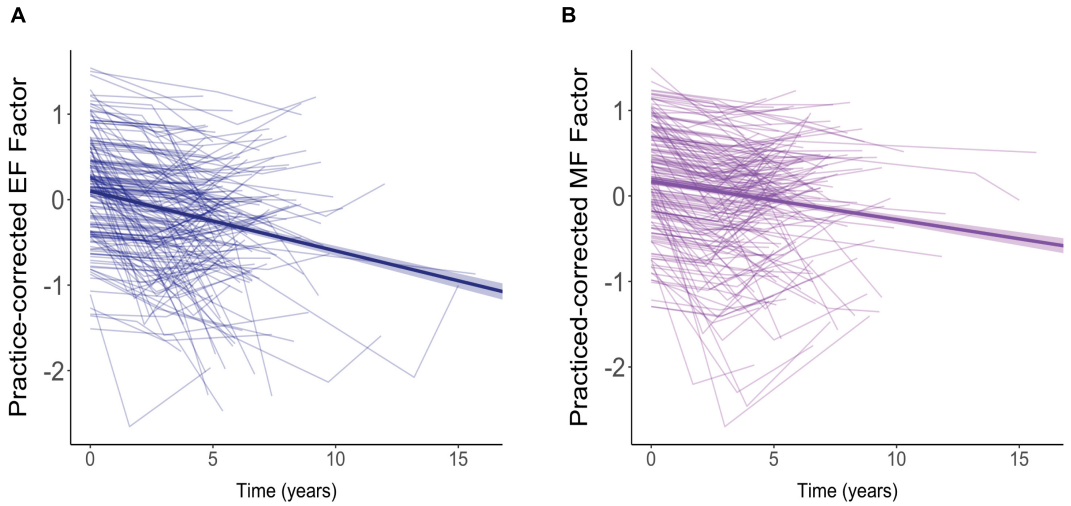


FIGURE 4 | Effects of aging on practice-corrected factor scores for **(A)** executive function (EF) and **(B)** memory function (MF). Each participant's scores across the three test sessions are connected by a thin, light-colored line. The dark blue (EF) and purple (MF) lines reflect the overall trend in factor scores with each additional year of age. The colored ribbon around these lines is the 95% confidence interval.

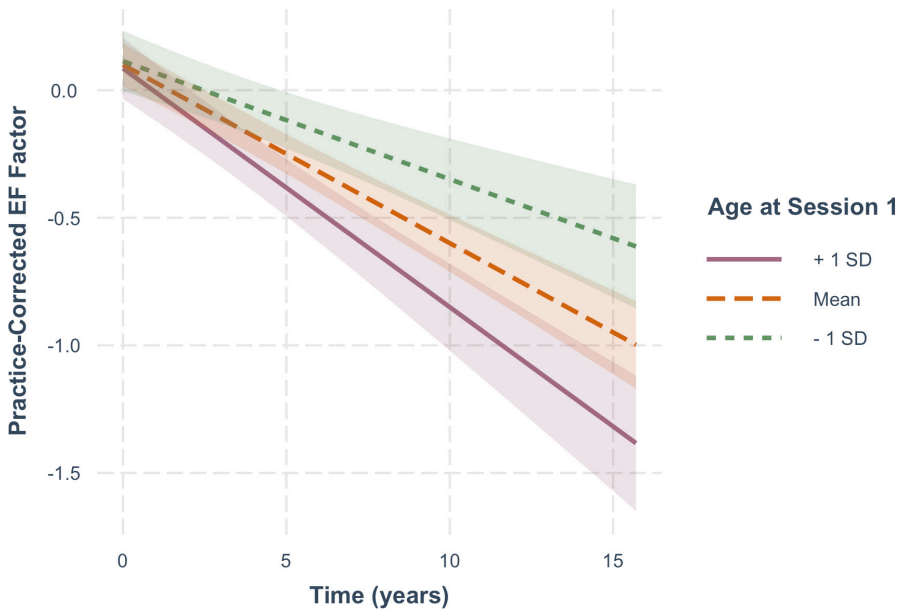


FIGURE 5 | The moderating effect of baseline age on the rate of change in practice-corrected EF scores with each year of aging. See **Figure 2** for details.

was greater in individuals who were older on average at baseline, accounting for the interaction.

For practice-corrected MF scores, decline over time was significantly moderated by baseline age (β = −0.003, *SE* = 0.001, $p = 0.009$), and also by FSIQ (β = 0.001, *SE* = 0.0005, $p = 0.022$), but not by sex (β = −0.016, *SE* = 0.013, $p = 0.217$). As shown in **Figure 6**, practice-corrected MF scores significantly decreased over time regardless of baseline age (1 *SD* older than the mean: β = −0.057, *SE* = 0.009, $p < 0.001$; mean age: β = −0.040, *SE* = 0.007, $p < 0.001$; 1 *SD* younger than the mean: β = −0.024, *SE* = 0.009, $p = 0.007$), but the rate of decline was greater in individuals who were older on average at baseline. Similarly, as shown in **Figure 7**, practice-corrected MF scores significantly decreased over time regardless of FSIQ (1 *SD* above the mean: β = −0.027, *SE* = 0.010, $p = 0.007$; mean FSIQ: β = −0.041, *SE* = 0.007, $p < 0.001$; 1 *SD* below the mean: β = −0.055, *SE* = 0.009, $p < 0.001$), but the rate of decline was slower in individuals who had higher FSIQs. Although FSIQ did not significantly moderate change in practice-corrected EF scores, this finding is shown in **Figure 8** for comparison purposes.

Similar to the age-corrected scores, there were cross-sectional effects of FSIQ and sex at baseline. Higher FSIQ scores were

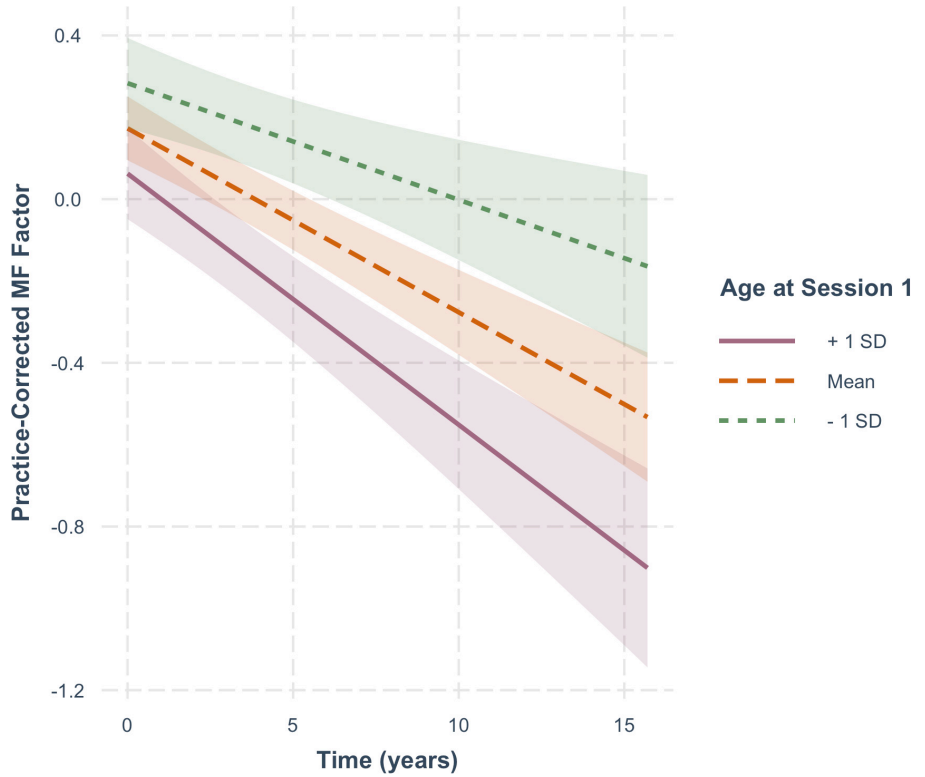


FIGURE 6 | The moderating effect of baseline age on rate of change in practice-corrected MF scores with each year of aging. See **Figure 2** for details.

associated with higher baseline scores on both cognitive measures (EF: β = 0.023, *SE* = 0.003, $p < 0.001$; MF: β = 0.019, *SE* = 0.003, $p < 0.001$). Men had higher baseline EF scores (β = 0.196, *SE* = 0.093, $p = 0.037$) and women had higher baseline MF scores (β = −0.615, *SE* = 0.087, $p < 0.001$). However, for these analyses of age-uncorrected scores, there was no significant effect of age on baseline EF scores (β = −0.003, *SE* = 0.008, $p = 0.742$), but there was an effect on MF scores (β = −0.023, *SE* = 0.008, $p = 0.006$), such that individuals who were older at baseline had lower baseline MF scores. This is consistent with the age correction being greater for MF than for EF scores.

DISCUSSION

In the present longitudinal study, we found that in a group of normally-aging older adults (65+), significant age-corrected retest effects (across three sessions in approximately six years) occurred in episodic memory but not in executive function. On the other hand, normal aging independent of practice effects, appeared to have similar effects on the two cognitive domains, resulting in significant declines in both cognitive functions. In general, young-older adults did better than old-older adults, showing greater practice effects and slower rates of decline with age. Interestingly, although full-scale IQ was associated with higher levels of performance at baseline for both cognitive factors, higher IQs did not enhance practice effects but were associated with a slower rate of age-related decline in memory; they did not significantly moderate the decline in executive function.

Practice Effects

These results, as a whole, seem to make it clear that practice effects do not occur equally across all cognitive domains; that is, there is no general cognitive practice effect. Previous studies have reported similar findings. For example, Hassenstab et al. (2015) found practice effects in episodic memory, but not in several other cognitive domains including executive function (using tests similar to ours) (see also Wilson et al., 2006, However, Gross et al. (2015) did report practice effects in executive function, but with tests that were mostly non-overlapping with the ones used here. Elman et al. (2018), also using a different set of tests, found smaller practice effects in executive function than in episodic memory even in a younger group of adults (aged 50-60), but no practice effects in executive function when baseline cognitive ability was controlled. Together, these findings suggest that practice effects may be domain-specific, or possibly process-specific, occurring reliably in episodic memory but not in executive function, at least in the executive functions that were captured by our EF factor. Our findings also indicate that, when testing memory, one must account for practice effects because they can mask the effects of aging even at long delays (see also Rönnlund et al., 2005; Elman et al., 2018). As seen in **Table 1**, *z*-scores on the age-independent memory composite increase from 0.18 to 0.36 to 0.46 (*z*-scores) across the three tests, showing

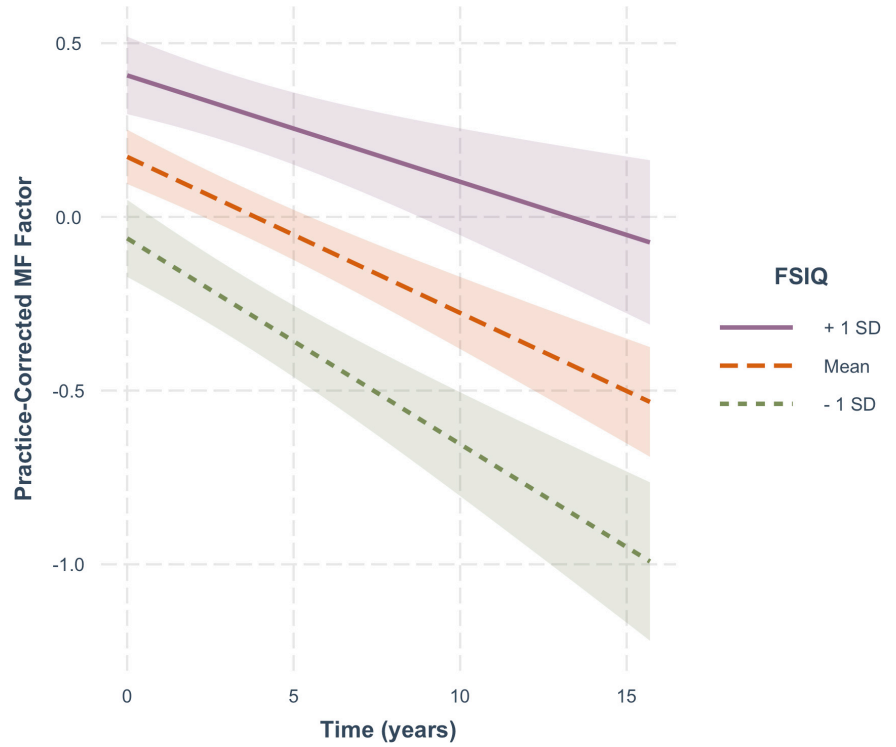


FIGURE 7 | The moderating effect of full scale IQ (FSIQ) on the rate of change in practice-corrected MF scores with each year of aging. The solid line represents those who on average have FSIQs 1 SD above the mean, the large-dashed line shows the group at the mean, and the small-dashed line portrays those whose FSIQ scores were 1 SD below the mean. The colored ribbon around each line is the 95% confidence interval.

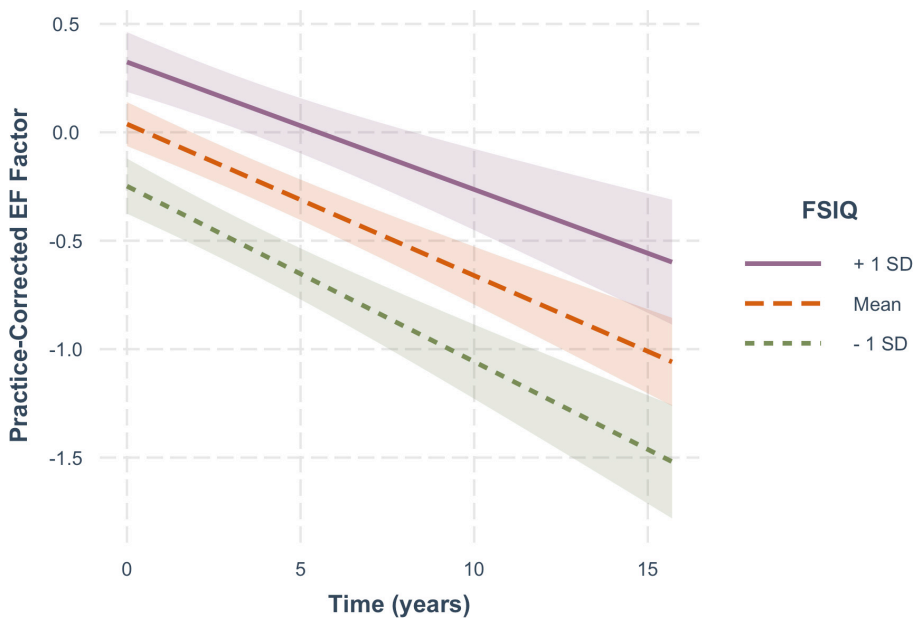


FIGURE 8 | The non-significant moderating effect of FSIQ on the rate of change in practice-corrected EF scores with each year of aging. See **Figure 7** for details.

robust beneficial effects of repeated testing. If variance due to age had not been removed from those scores, those scores would have been 0.24, 0.29, and 0.28, and conclusions might have been that episodic memory seems to hold up well with normal aging. Nevertheless, our older people did show declines with age, once practice effects were removed.

So why did retesting improve memory but not executive function? For episodic memory, two possible answers to this question have generally been suggested (Goldberg et al., 2015): (a) People remember some of the actual stimuli from a first test and are therefore able to learn more, and (b) people develop memory strategies during the prior experience, which could later be employed to enhance memory further. Given the variety of memory tests that made up the composite measure, it seems unlikely that a common strategy or multiple strategies would have been learned during a single testing session. In the present study, memory continued to grow across two successive retests, suggesting that people were accessing the same memory representation and strengthening it on each occasion. We know that in the short term, repetition strengthens a memory trace. In the long term, retesting might enhance retrieval of a memory by presenting partial cues. New information might then be added to and strengthen the trace, which is then reconsolidated. In our sample, although the original memory traces may have weakened over time, they appeared still to be available and accessible when good cues were provided at retest. This explanation for retest effects fits well with the assumption that consolidation was the common memory process across the five tests that comprised the memory composite.

For executive function, although repetition might have allowed one to access the prior experience, it might not have helped one to perform the executive function tasks more efficiently. The tasks that comprised the EF factor in this sample all required attentional focus in the presence of interference, such as those involved in most working memory tasks. These kinds of tasks and processes lend themselves less well to the benefits of practice; gains tend to be short-term and task-specific, and require long hours of training (see Baddeley et al., 2015). Thus, it is not surprising that our EF factor did not improve across just two additional test sessions over several years.

Neither sex nor FSIQ influenced practice effects in either cognitive domain, suggesting that the practice effects that occurred in memory, may be at least partly automatic in normally-aging individuals. Whether you are a young-older person or amongst the oldest-old, intellectually gifted or less so, male or female, practice will enhance episodic memory. Our results did suggest, however, that improvements associated with retesting in the memory domain were smaller in the oldest, older adults. Interpretation of this finding, however, is not straightforward. It may reflect a decline in some automatic processes that are activated during retesting. For example, although cues from current tests may activate memory representations of prior sessions in older adults, the activation process might be slower or less complete at older ages. On the other hand, the representations themselves might be weaker in older adults, leading to a smaller increase over tests. It should be noted, however, that the baseline levels of performance at test session 1 differed across age groups, with the older adults having higher scores, and all age groups performing approximately equivalently by the third session. This suggests that ceiling effects might have reduced performance over time for those with higher levels of performance at baseline, in this case, the oldest group. The use of composite measures, however, limits ceiling effects and the composite scores did not appear to approach the ceiling. So overall, with respect to memory function, we conclude that normally-aging older adults of all ages show significant benefits of practice, although benefits may be smaller at oldest ages.

For executive function, there was a small but non-significant decrease in performance across test sessions indicating no effects of practice. Here, performance at baseline was also significantly higher for the old-older adults, and the scores also converged across age groups by the third test. Whereas the oldest group showed a significant decrease in performance across test sessions, the youngest group showed a non-significant increase. It is unclear why the oldest group's performance would have declined across repeated tests. There may be another variable associated with re-testing that negatively impacts performance on our executive function tests. Alternatively, these results may reflect a regression to the mean. The bottom line, however, is that we did not find any significant benefits of retesting for executive function.

What might seem rather anomalous in these findings is that in both cognitive composites after age correction, the old-older adults, in general, were performing at a higher level at baseline than young-older adults. As noted earlier, however, individuals who dropped out of the study for various reasons or were removed for failing to meet inclusion criteria, tended to be older and had lower levels of cognitive performance. This resulted in a sample that was younger than the reference group on which their scores were based, leading to more below average composite scores amongst the young-older adults (i.e., a negative age correction) and higher composite scores amongst the oldest-old (i.e., a positive age correction). In the age-uncorrected data, the oldest adults showed the expected lower levels of performance, particularly in memory (see **Figure 6**, Time 0).

Aging Effects

For these analyses, measures represent age-uncorrected performance levels. In **Figure 5**, one can see that there are no significant cross-sectional effects of age at baseline for executive function, but a significant effect of baseline age on memory function (**Figure 6**). This differential effect of age on the two cognitive composites accounts for the greater age-correction in memory than in executive function (see **Table 1**).

Aging effects, namely change in performance over time/years without any benefit from retests, are clearly evident in both cognitive domains. In addition, baseline age moderated both functions similarly, with the old-older people showing steeper declines over time than the young-older people. This finding is consistent with the notion that there might be a general aging-related factor common to the two domains (cf., Wilson et al., 2002; Salthouse, 2003).

At the same time, however, FSIQ, which was associated with baseline levels of performance for both cognitive functions, had no significant association with aging-related decline in executive function, but a significant moderating effect on memory function, such that those with higher IQs exhibited a slower decline in memory than those with lower IQs. This suggests an age-related process or function that differs across the two cognitive domains. Most longitudinal studies of aging have not included

IQ as an individual difference variable although several have included education, with mixed results. We decided to include FSIQ, rather than education, primarily because of the different educational opportunities available to people across this wide age range, such that less education in our oldest old might not necessarily translate into lower intellectual function. We expected that IQ would incorporate not only acquired knowledge and skills gained in an educational context, but also a broader range of experiences and abilities acquired over a lifetime. In addition, education has often not shown any influence on age-related decline in memory (e.g., Zahodne et al., 2011; Wilson et al., 2019) or other cognitive functions (for review, see Seblova et al., 2020). In the aging literature generally, both education and IQ have been used as proxies for what has been called cognitive reserve (see Stern, 2007, 2009) and it is in this context that we will discuss the possible impact of IQ in the present study.

Here there are two related questions to be considered: Why is FSIQ associated with performance at baseline in both cognitive domains, and why does it moderate age-related decline only in memory? Reserve theory would suggest that baseline performance levels in both domains are related to brain reserve, which is established through the development of a structurally "better" brain (e.g., greater volume or connectivity) resulting from more varied life activities and experiences, and is reflected in the IQ measures. Brain reserve may benefit cognitive functions more broadly, as evidenced by the higher levels of performance at baseline in both executive and memory function for those with higher IQs. Although brain reserve is considered to be a relatively fixed entity at any one time, it also needs to be maintained over time presumably by continuing engagement in life's activities (see Stern et al., 2020, 2022 for further elaboration). Cognitive reserve, however, refers to a more flexible and dynamic ability to adapt one's cognitive processing in the light of declining brain networks. Thus, in the present study, high IQ at baseline may reflect greater brain reserve, which is supporting higher levels of memory and executive function at baseline, whereas the ability to moderate cognitive decline over time may reflect cognitive reserve, which may be domain- or process-specific. In memory, for example, older people tend to be more reliant on cues than younger people to retrieve episodic memories. Those adults with greater cognitive reserve may make more effective use of cues at retrieval, and therefore be more likely to reactivate a fading memory trace. On the other hand, the executive control processes associated with working memory, namely attentional focus under conditions of interference, may be less adaptable, and so less responsive to cognitive reserve. Note (see **Figure 8**) that there was a smaller but non-significant effect of FSIQ on age-related changes in executive function.

Overall, these results suggest that there may be both a common factor related to age-related declines in both cognitive functions, but also domain-specific factor(s) that might be differentially effective for different cognitive functions or processes.

Implications

The present results indicate that both episodic memory and the executive functions associated with working memory decline with age. They also suggest that episodic memory may be more amenable to intervention than executive function in normally-aging older adults; practice improves memory and cognitive reserve helps to slow its decline. However, our sample included only people who were determined to be aging "normally," and therefore does not speak to whether practice or cognitive reserve could be recruited to help those people with mild cognitive impairment or dementia. Prior studies that have included those with cognitive impairments are inconsistent in this respect with some studies showing improvements across retests (e.g., Gross et al., 2015) and others showing minimal or no effects (e.g., Hassenstab et al., 2015). In the present study of normally-aging older people, however, improvements in memory seemed to be available to even the oldest old, although perhaps to a somewhat lesser degree with increasing age.

From both a research and clinical perspective, the present findings have a number of implications. The results are based on a sample of people that are cognitively normal across all three testing sessions. They do not include people who have given any indication of underlying pathology that might affect cognitive function at any time over the years, although clearly, in the absence of any brain measures, we cannot rule that out. The sample, however, is relatively high-functioning with only a small number of individuals with IQs below 100. Although IQ is not reported in many studies, several have noted education levels of 16 years, comparable to our study (e.g., Wilson et al., 2006; Salthouse, 2010; Armstrong et al., 2020). Thus we do not think our sample is unique in that respect. It is, however, possible that although there was no effect of IQ on practice effects, those with still lower IQs might not show such benefits. Nevertheless, we think that the sample in this study is a good representation of normal cognitive aging in a community-based sample against which other comparable samples may be compared. We also feel confident in concluding that people who are aging normally should show practice effects on memory tests, but not necessarily on tests that require working memory or tax attentional resources. Failure to show retest effects on memory tests should therefore be considered a possible indicator of abnormal aging, which should be evaluated further.

Clinically, when assessing an older person on more than one occasion, especially in memory and even at long intervals, one needs to be aware that simply repeating the tests may confer some advantage and so scores may overestimate ability. One might want to choose different memory tests or materials at retest to offset, at least partly, the effects of practice, although if strategies were learned at initial testing, they might still provide some benefit. Accounting for practice effects may be particularly important for accurate diagnosis of mild cognitive impairment, particularly amnestic MCI. Eliminating the effects of retests may enable earlier diagnosis and intervention, which may prevent or slow the progression of the disease (see Elman et al., 2018; Sanderson-Cimino et al., 2022). Acknowledging possible effects of retesting might also be important in such things as clinical trials designed to evaluate the effects of a drug, for example (e.g., Goldberg et al., 2015). At the same time, if one is interested in interventions with real-world applications for normally-aging older adults, using a repeated testing procedure is a well-known strategy for enhancing memory

over time (Roediger and Karpicke, 2006). Attempted retrieval of previously learned information has also been shown to improve memory and enhance learning of new information in people with memory impairments including those with Alzheimer's disease (e.g., Pastötter and Bäuml, 2014). Retention intervals in these studies, however, are usually quite short (i.e., one month).There are thus both positive and negative effects of retesting: In longitudinal studies of aging, retesting may mask age-related declines in memory, leading to missed diagnoses of MCI, but in clinical interventions, retrieval practice may enhance memory in everyday life.

Finally, we would like to re-emphasize that the failure to find effects of practice or cognitive reserve in executive function very likely depends on the specific tests and processes. Executive function tasks rely on multiple processes, and although there may be a common factor across tests, there are clearly several different executive control processes grouped under the banner of executive function (e.g., Miyake et al., 2000; Glisky et al., 2021). Some of these may be modifiable by cognitive reserve or susceptible to practice, others may not. Looking at different types of executive functions longitudinally in an aging population would be an important future endeavor, which could identify more specifically the kinds of processes that are most amenable to modification. These findings also support the benefits of using composite measures made up of tests that might differ in many ways but share a common process. Being able to identify specific processes that are affected by aging, rather than focusing just at the domain level, could further enhance our understanding of aging and suggest interventions most likely to succeed.

Strengths and Limitations

One of the major strengths of this study, as already noted, was the high probability that our sample included only older adults who were aging normally with respect to their cognitive function. This reduced the likelihood that any negative outcomes that we observed might be attributable to incipient pathology. At the same time, however, our sample was quite high functioning and may not be representative of the population in general. Second, as suggested and incorporated by many others, we used composite scores to reduce variability and error, but in our study (as in some others), the tests comprising the composites were chosen to reflect a common process determined through factor analysis. This allowed us to go beyond what many have said before about what cognitive domains are or are not affected by aging, and to begin identification of specific processes. Third, we believe that we have introduced a relatively novel way of separating practice and aging effects within an individual across repeated tests. Many studies have looked at practice effects across individuals, by comparing Time 1 performance in those who completed only Time 1 to those from the same cohort at Time 2, but this comparison is still between-persons and could be affected by other individual differences. Finally, we think that our results showing robust within-person practice effects in memory and no practice effects in our measure of executive function, make a strong case for concluding that not all cognitive functions show improvements with practice or retesting, and leaves room for many more studies to explore this issue at the level of processes. The findings with respect to aging also leave open the possibility that there may be (a) a common age-related factor that affects all cognitive processes, for example, global changes in the brain, (b) a common domain-related factor that affects all tests within a domain, or (c) process-specific factors within domains, dependent on more specific brain regions.

Limitations of our study include a relatively small sample size. In longitudinal studies that rely on community-based older adult volunteers who need to be available for several years, there are always many dropouts for a variety of reasons. In our case, to ensure that our sample continued to age normally, we also excluded people who had or developed psychiatric or neurological conditions that might affect cognitive function. Our sample size therefore limited to some degree the kinds of analyses that we could do and our ability to explore additional factors. Another limitation of our work is that we did not have any direct measures of brain integrity or function, which might support our cognitive findings. Although we suggested that the common factor among our memory tests most likely reflected consolidation dependent on medial temporal lobe regions, and our executive function tests depended on prefrontal brain regions associated with working memory, we could not determine that from our study, and certainly we could not be more specific. The recent advances in neuroimaging, however, which have begun to relate longitudinal changes in cognitive functions to corresponding changes in different brain regions (e.g., Persson et al., 2014; Armstrong et al., 2020; Gavett et al., 2021), will continue to lead to new ideas and discoveries that will add considerably to our growing understanding of both normal and pathological aging.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Arizona Human Subjects Protection Program. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EG: conceptualization, methodology, supervision, and writing – original draft, review and editing. CW: data curation, investigation, and project administration. KM: data curation and formal analysis. MG: formal analysis and supervision, methodology, visualization, and writing – original draft, review and editing. Based on contributor roles taxonomy (CRediT). All authors contributed to the article and approved the submitted version.

FUNDING

We gratefully acknowledge support from the National Institute on Aging (AG014792), Arizona Alzheimer's Consortium, Arizona Department of Health Services, and the McKnight Brain Research Foundation.

ACKNOWLEDGMENTS

Thanks to the many research assistants and graduate students who contributed to data collection and database maintenance over the years and to all of the older adults from the community who volunteered to be a part of this project.

REFERENCES

Armstrong, N. M., An, Y., Shin, J. J., Williams, O. A., Doshi, J., Erus, G., et al. (2020). Associations between cognitive and brain volume changes in cognitively normal older adults. *Neuroimage* 223:117289. doi: 10.1016/j.neuroimage.2020.117289

Baddeley, A., Eysenck, M. W., and Anderson, M. C. (2015). *Memory*, 2nd Edn. London: Psychology Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Craik, F. I. M., Morris, L. W., Morris, R. G., and Loewen, E. R. (1990). Relations between source amnesia and frontal lobe functioning in older adults. *Psychol. Aging* 5, 148–151. doi: 10.1037/0882-7974.5.1.148

Delis, D. C., Kramer, J., Kaplan, E., and Ober, B. A. (1987). *The California Verbal Learning Test*. San Antonio, TX: Psychological Corporation.

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., et al. (2018). Underdiagnosis of mild cognitive impairment: a consequence of ignoring practice effects. *Alzheimers Dement. (dadm)* 10, 372–381. doi: 10.1016/j.dadm.2018.04.003

Ferrer, E., Salthouse, T. A., McArdle, J. J., Stewart, W. F., and Schwartz, B. S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychol. Aging* 20, 412–422. doi: 10.1037/0882-7974.20.3.412

Ferrer, E., Salthouse, T. A., Stewart, W. F., and Schwartz, B. S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychol. Aging* 19, 243–259. doi: 10.1037/0882-7974.19.2.243

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/-22-3956(75)90026-6

Gavett, B. E., Fletcher, E., Widaman, K. F., Farias, S. T., DeCarli, C., and Mungas, D. (2021). The latent factor structure underlying regional brain volume change and its relation to cognitive change in older adults. *Neuropsychology* 35, 643–655. doi: 10.1037/neu0000761

Glisky, E. L., Alexander, G. E., Hou, M., Kawa, K., Woolverton, C. B., Zigman, E. K., et al. (2021). Differences between young and older adults in unity and diversity of executive functions. *Aging Neuorpsychol. Cogn.* 28, 829–854. doi: 10.1080/13825585.2020.1830936

Glisky, E. L., and Kong, L. L. (2008). Do young and older adults rely on different processes in source memory tasks? A neuropsychological study. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 809–822. doi: 10.1037/0278-7393.34.4.809

Glisky, E. L., Polster, M. R., and Routhieaux, B. C. (1995). Double dissociation between item and source memory. *Neuropsychology* 9, 229–235. doi: 10.1037/0894-4105.9.2.229

Glisky, E. L., Rubin, S. R., and Davidson, P. S. R. (2001). Source memory in older adults: an encoding or retrieval problem? *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1131–1146. doi: 10.1037/0278-7393.27.5.1131

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement. (dadm)* 1, 103–111. doi: 10.1016/j.dadm.2014.11.003

Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., et al. (2015). Predictors of retest effects in a longitudinal study of cognitive aging in a diverse community-based sample. *J. Int. Neuropsychol. Soc.* 21, 506–518. doi: 10.1017/S1355617715000508

Hart, R. P., Kwentus, J. A., Wade, J. B., and Taylor, J. R. (1988). Modified Wisconsin card sorting test in elderly normal, depressed, and demented patients. *Clin. Neuropsychol.* 2, 49–56. doi: 10.1080/13854048808520085

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effect in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Janowsky, J. S., Shimamura, A. P., and Squire, L. R. (1989). Source memory impairment in patients with frontal lobe lesions. *Neuropsychologia* 27, 1043–1056. doi: 10.1016/0028-3932(89)90184-X

Kramer, J. H., Mungas, D., Reed, B. R., Wetzel, M. E., Burnett, M. M., Miller, B. L., et al. (2007). Longitudinal MRI and cognitive change in healthy elderly. *Neuropsychology* 21, 412–418. doi: 10.1037/0894-4105.21.4.412

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmer test package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the mayo clinic study of aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

McCabe, D. P., Roediger, H. L. III, McDaniel, M. A., and Hambrick, D. Z. (2010). The relationship between working memory capacity and executive function: evidence for a common executive attention construct. *Neuropsychology* 24, 222–243. doi: 10.1037/a0017619

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Pastötter, B., and Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Front. Psychol.* 5:286. doi: 10.3389/fpsyg.2014.00286

Persson, J., Pudas, S., Nilsson, L.-G., and Nyberg, L. (2014). Longitudinal assessment of default-mode brain function in aging. *Neurobiol. Aging* 35, 2107–2117. doi: 10.1016/j.neurobiolaging.2014.03.012

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabbitt, P., Diggle, P., Smith, D., Holland, F., and McInnes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia* 39, 532–543. doi: 10.1016/S0028-3932(00)0009-3

Roediger, H. L. III, and Karpicke, J. D. (2006). Test-enhanced learning:taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280

Rönnlund, M., Nyberg, L., Bäckman, L., and 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, 3–18. doi: 10.1037/0882-7974.20.1.3

Salthouse, T. A. (2003). Memory aging form 18 to 80. *Alzheimer Dis. Assoc. Disord.* 17, 162–167. doi: 10.1097/00002093-200307000-00008

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology* 24, 563–572. doi: 10.1037/a0019026

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., et al. (2022). Cognitive practice effects delay diagnosis of MCI: implications for clinical trials. *Alzheimers Dement. (dadm)* 8:e12228. doi: 10.1002/trc2.12228

Schacter, D. L., Harbluk, J. L., and McLachlan, D. R. (1984). Retrieval without recollection: an experimental analysis of source amnesia. *J. Verbal Learn. Verbal Behav.* 23, 593–611. doi: 10.1016/S0022-5371(84)90373-6

Seblova, D., Berggren, R., and Lövdén, M. (2020). Education and age-related decline in cognitive performance: systematic review and meta-analysis of longitudinal cohort studies. *Ageing Res. Rev.* 58:101005. doi: 10.1016/j.arr.2019.101005

Spreen, O., and Benton, A. L. (1977). *Neurosensory Center Comprehensive Examination for Aphasia* (rev. ed.). Victoria, BC: University of Victoria of Victoria Neuropsychology Laboratory.

Stern, Y. (2009). Cognitive reserve. *Neuropsychologia* 47, 2015–2028. doi: 10.1016/j.neuropsychologia.2009.03.004

Stern, Y., Albert, M., Barnes, C., Cabeza, R., Pascual-Leone, A., and Rapp, P. (2022). "Framework for terms used in research of reserve and resilience," in *Collaboratory on Research Definitions for Reserve and Resilience in Cognitive Aging and Dementia*. Available online at: https://reserveandresilience.com/framework (accessed February 18, 2022).

Stern, Y., Arenaza-Urquijo, E. M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., et al. (2020). Whitepaper: defining and investigating cognitive reserve, brain reserve, and brain maintenance. *Alzheimers Dement*. 16, 1305–1311. doi: 10.1016/j.jalz.2018.07.219

Stern, Y. (ed.) (2007). *Cognitive Reserve: Theory and Applications*. New York, NY: Taylor & Francis.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale—Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale—III*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale—III*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio, TX: Psychological Corporation.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wilson, R. S., Beckett, L. A., Barnes, L. L., Schneider, J. A., Bach, J., Evans, D. A., et al. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychol. Aging* 17, 179–193. doi: 10.1037/0882-7974.17.2.179

Wilson, R. S., Li, Y., Bienias, J. L., and Bennett, D. A. (2006). Cognitive decline in old age: separating retest effects from the effects of growing older. *Psychol. Aging* 21, 774–789. doi: 10.1037/0882-7974.21.4.774

Wilson, R. S., Yu, L., Lamar, M., Schneider, J. A., Boyle, P. A., and Bennett, D. A. (2019). Education and cognitive reserve in old age. *Neurology* 92, e1041–e1050. doi: 10.1212/WNL.0000000000007036

Zahodne, L. B., Glymour, M. M., Sparks, C., Bontempo, D., Dixon, R. I., MacDonald, S. W., et al. (2011). Education does not slow cognitive decline with aging: 12-year evidence from the victoria longitudinal study. *J. Int. Neuropsychol. Soc.* 17, 1039–1046. doi: 10.1017/S1355617711001044

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fnagi.2022.847315

Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments

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

[1] University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego State University, San Diego, CA, United States, [2] Center for Behavior Genetics of Aging, University of California, San Diego, San Diego, CA, United States, [3] Department of Psychiatry, School of Medicine, University of California, San Diego, San Diego, CA, United States, [4] Department of Family Medicine and Public Health, University of California, San Diego, San Diego, CA, United States, [5] Sam and Rose Stein Institute for Research on Aging, University of California, San Diego, San Diego, CA, United States, [6] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MA, United States, [7] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States, [8] Psychology Service, VA San Diego Healthcare System, San Diego, CA, United States, [9] Research Service, VA San Diego Healthcare System, San Diego, CA, United States, [10] Rehabilitation Institute of Washington, Seattle, WA, United States, [11] Center of Excellence for Stress and Mental Health, Veterans Affairs San Diego Healthcare System, San Diego, CA, United States, [12] Department of Psychological and Brain Sciences, Boston University, Boston, MA, United States

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 twofold 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

OPEN ACCESS

Edited by:
Claudia Jacova,
Pacific University, United States

Reviewed by:
Maria Josefsson,
Umeå University, Sweden
Erika J. Laukka,
Karolinska Institutet (KI), Sweden

***Correspondence:**
Mark Sanderson-Cimino
mesander@health.ucsd.edu

†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

Specialty section:
This article was submitted to Alzheimer's Disease and Related Dementias, a section of the journal Frontiers in Aging Neuroscience

Received: 02 January 2022
Accepted: 21 March 2022
Published: 25 April 2022

Citation:
Sanderson-Cimino M, Elman JA, Tu XM, Gross AL, Panizzon MS, Gustavson DE, Bondi MW, Edmonds EC, Eppig JS, Franz CE, Jak AJ, Lyons MJ, Thomas KR, Williams ME and Kremen WS (2022) Practice Effects in Mild Cognitive Impairment Increase Reversion Rates and Delay Detection of New Impairments. Front. Aging Neurosci. 14:847315. doi: 10.3389/fnagi.2022.847315

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.

Keywords: practice effects, cognitive aging, mild cognitive impairment, Alzheimer's disease, biomarkers, dementia progression

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 (Manly et al., 2008; Albert et al., 2011). 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) (Manly et al., 2008; Albert et al., 2011; Eppig et al., 2020; 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 and 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 (Edmonds et al., 2015a; 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% (Cummings et al., 2014; Anand et al., 2017). 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 (Sperling R. et al., 2014; Sperling R. A. et al., 2014; Canevelli et al., 2016; Anand et al., 2017; Alexander et al., 2021). As noted by Canevelli et al. (2016), at least 274 randomized controlled trials were recruiting MCI subjects in 2016. As such, accurate diagnoses of earlier disease stages are necessary to further the treatment of AD (Edmonds et al., 2018; Veitch et al., 2019; Eppig et al., 2020).

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 (2016) and Pandya et al. (2016) also suggested that reducing reversion rates should be an essential goal of future MCI methodology studies. Canevelli et al. (2016) and Pandya et al. (2016) 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. Malek-Ahmadi (2016) 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. 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 Mild Cognitive Impairment

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 (Ronnlund et al., 2005; Gross et al., 2015; Elman et al., 2018; 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., 2020).

The MCI classification methods, particularly in research, almost always rely on use of cut-off scores to define cognitive impairment (Winblad et al., 2004; Jak et al., 2009). 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 was 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[1]. 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.

Mild cognitive impairment was diagnosed using the Jak-Bondi approach (Jak et al., 2009; Bondi et al., 2014; Edmonds et al., 2018). 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., 2018).

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[2]: Aβ+: Aβ < 977 pg/mL; p-tau+: p-tau > 21.8 pg/mL; t-tau+: t-tau > 270 pg/mL (Hansson et al., 2018; Elman et al., 2020). 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 0.5 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.

[1] adni.loni.usc.edu

[2] http://adni.loni.usc.edu/methods

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. (2012) suggested that the most suitable approach is to utilize replacement participants (Rönnlund and Nilsson, 2006). To our knowledge, the replacement-participant approach has only been utilized in two samples (Ronnlund et al., 2005; Elman et al., 2018). 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 (Ronnlund et al., 2005; Elman et al., 2018).

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 et al., 2011, 2014; Duff, 2014; Duff and Hammers, 2020). 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., 2020). 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) (Ho et al., 2018). Additional *t*-tests and chi-squared tests ensured that returnees and pseudo-replacements were matched at a group level ($ps > 0.8$). 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 and replacements) varied but was always greater than 80 participants.

The equations below were used to calculate the PE:

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

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

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

Where Returnees_{T2} represents the mean score of the returnee sample at their second assessment, $\text{Pseudo-replacements}_{T1}$ 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 $\text{Proportional Baseline}_{T1}$ 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 (Ronnlund et al., 2005; Elman et al., 2018). 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, 16–36, 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 $d = 0.06–0.26$). 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, and 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: 6.5–13.3%), but this equated to only a small change in the number of participants (ns < 5).

TABLE 1A | 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.

TABLE 1B | 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	−0.01 [−0.13, 0.16]	0.09 [−0.10, 0.43]	−0.40 [−1.57, 0.89]	−2.31 [−6.64, 2.27]	0.11 [−0.14, 0.33]	0.43 [0.15, 0.72]
Practice effect	0.80 [−0.33, 3.08]	0.89 [−0.41, 3.33]	−1.64 [−5.65, 2.41]	−4.36 [−19.16, 9.57]	0.63 [−0.21, 1.53]	NA
Cohen's *d*	0.26	0.20	−0.07	−0.06	0.14	NA

Groups are based on the average performance across all 5,000 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 2 | 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; $p < 0.001$	5.1; $p = 0.02$	7.5; $p = 0.006$	22.0; $P < 0.001$	3.2; $p = 0.07$	3.2; $p = 0.07$	21.0; $p < 0.001$

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.

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 so 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 ($W = 1703$; $p = 0.177$). 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 ($W = 1240$; $p = 0.017$).

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

TABLE 3 | 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; *p*-value	2.25; $p = 0.13$	11.13; $p < 0.001$	1.3; $p = 0.25$	20.0; $p < 0.001$	0.5; $p = 0.48$	0.5; $p = 0.48$

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 4 | 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 $p = 0.13$	18.1 $p < 0.001$	20.0 $p < 0.001$	0.5 $p = 0.48$	0.5 $p = 0.48$
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 5 | Progression to dementia.

	Full sample *N* = 159	Stable MCI		Reverters		False reverters *N* = 10
Months until DX		**Unadjusted *N* = 141**	**Adjusted *N* = 151**	**Unadjusted *N* = 18**	**Adjusted *N* = 8**	
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.

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 ($ps > 0.17$).

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%). **Figures 1, 2** displays hazard ratios and survival curves for all models. **Supplementary 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 (Sperling R. A. et al., 2014; Canevelli et al., 2016; Pandya et al., 2016; Alexander et al., 2021). 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 et al., 2011; Jack et al., 2013; Edmonds et al., 2015b; Veitch et al., 2019; Elman et al., 2020), 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 et al., 2017, 2018). 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 et al., 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 (Edmonds et al., 2015a; Eppig et al., 2017; Thomas et al., 2017, 2019). In a similar ADNI subsample, use of robust norms found a 1-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 normal 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 (Ronnlund et al., 2005; Elman et al., 2018). 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 (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., 2020) 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 and 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

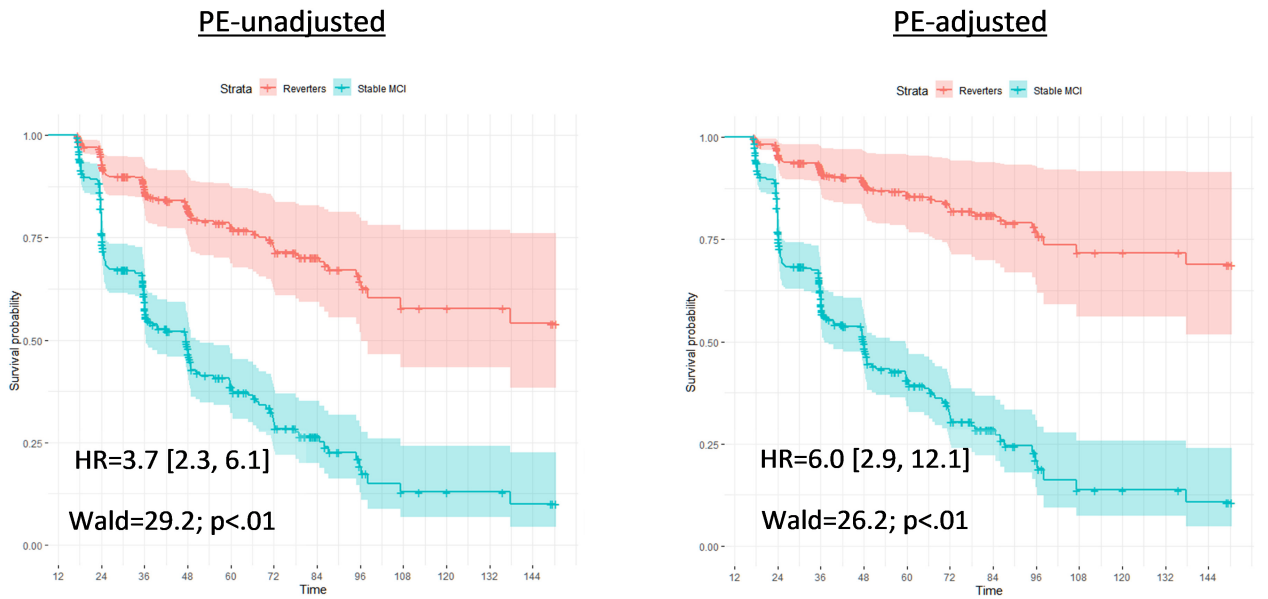


FIGURE 1 | Full Cox proportional models for time until first dementia diagnosis by PE-unadjusted and PE-adjusted 12-month diagnoses. Cox proportional hazard models compared progression to dementia between those who were classified with mild cognitive impairment at follow-up (Stable MCI) and those who 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 variable data (16 – months from 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 ratio (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.

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 non-significant. 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.

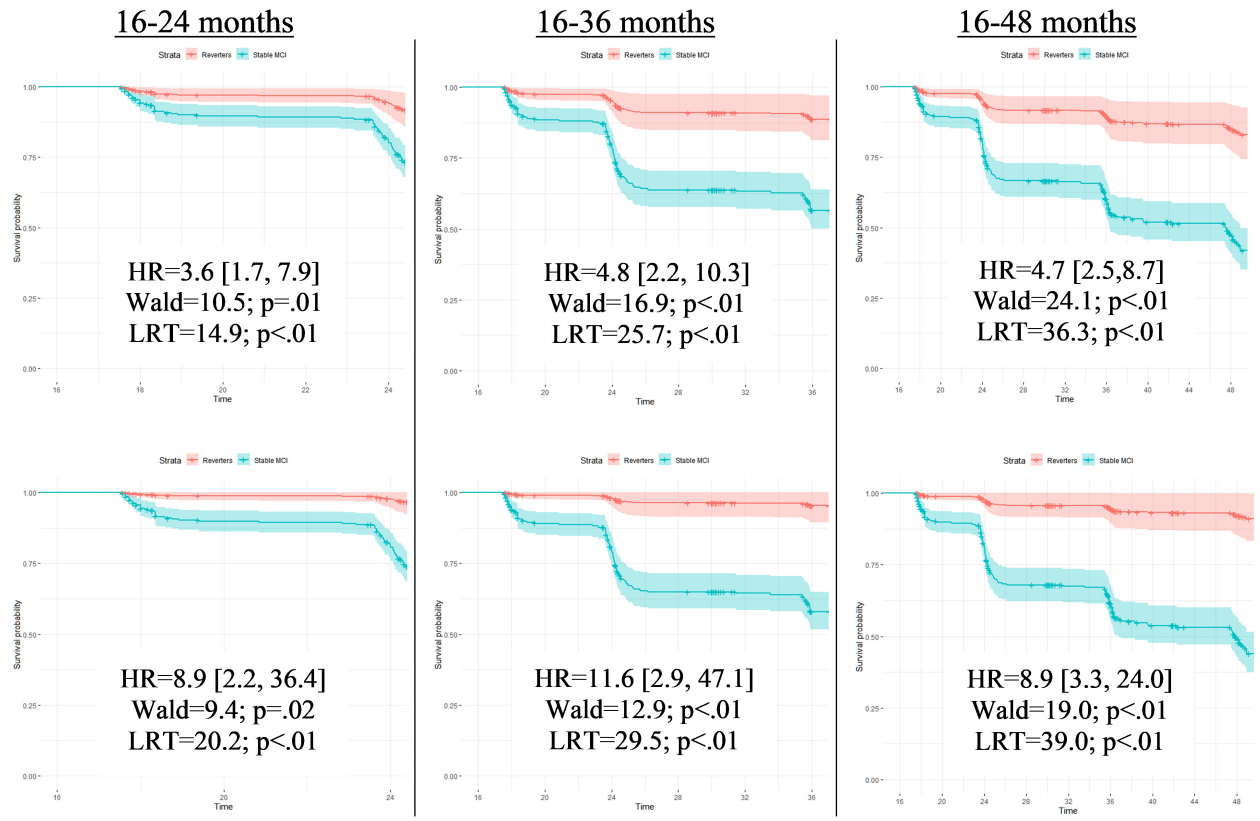


FIGURE 2 | Full Cox proportional models for time until first dementia diagnosis by PE-unadjusted and PE-adjusted 12-month diagnoses. Cox proportional hazard models compared progression to dementia between those who were classified with mild cognitive impairment at follow-up (Stable MCI) and those who reverted to cognitively normal (Reverters). 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 ratio (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.

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 (Ronnlund et al., 2005; Elman et al., 2018). 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., 2020). 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., 2020). As shown in the present study, the method can be adapted to large studies that did not include replacements in their original

TABLE 6A | Amyloid, total tau, and phosphorylated tau across classification groups.

	Full sample *N* = 226	Stable MCI		Reverters		False reverters *N* = 18
		Unadjusted *N* = 166	Adjusted *N* = 184	Unadjusted *N* = 60	Adjusted *N* = 42	
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 6B | Combined amyloid and tau positivity profiles.

	Full	Stable MCI		Reverters		False
	Sample	Unadjusted	Adjusted	Unadjusted	Adjusted	Reverters *n* = 18
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+).

design. However, building it into the original study design is clearly preferable.

CONCLUSION

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.

ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

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 was 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 Neuroimaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://adni.loni.usc.edu/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of California, San Diego. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MS-C and WK conceived the study. XT and AG provided guidance on statistical analysis. EE, MB, JSE, and KT made determination of MCI diagnoses. MS-C, WK, JAE, MP, and DG contributed to the practice effects methodology. WK, CF, ML, and MS-C obtained primary funding to support this work. All authors provided critical review and commentary on the manuscript.

FUNDING

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 United States National institute on Aging (MS-C: F31AG064834, WK, CF, and ML: R01 AG050595, CF and WK: P01 AG055367, WK: R01 AG022381, AG054002, and AG060470, CF: R01 AG059329, AG: K01 AG050699; MB: R01 AG049810, KT: 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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.847315/full#supplementary-material

REFERENCES

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (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. *Alzheimers Dement.* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Alexander, G. C., Emerson, S., and Kesselheim, A. S. (2021). Evaluation of aducanumab for Alzheimer disease: scientific evidence and regulatory review involving efficacy, safety, and futility. *JAMA* 325, 1717–1718. doi: 10.1001/jama.2021.3854

Anand, A., Patience, A. A., Sharma, N., and Khurana, N. (2017). The present and future of pharmacotherapy of Alzheimer's disease: a comprehensive review. *Eur. J. Pharmacol.* 815, 364–375. doi: 10.1016/j.ejphar.2017.09.043

Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., et al. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *J. Alzheimers Dis.* 42, 275–289. doi: 10.3233/JAD-140276

Braak, H., Thal, D. R., Ghebremedhin, E., and Del Tredici, K. (2011). Stages of the Pathologic Process in Alzheimer Disease: age Categories From 1 to 100 Years. *J. Neuropathol. Exp. Neurol.* 70, 960–969. doi: 10.1097/NEN.0b013e318232a379

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Canevelli, M., Grande, G., Lacorte, E., Quarchioni, E., Cesari, M., Mariani, C., et al. (2016). Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis. *J. Am. Med. Dir. Assoc.* 17, 943–948. doi: 10.1016/j.jamda.2016.06.020

Cummings, J. L., Morstorf, T., and Zhong, K. (2014). Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res. Ther.* 6, 1–7. doi: 10.1186/alzrt269

Dubois, B., Hampel, H., Feldman, H. H., Scheltens, P., Aisen, P., Andrieu, S., et al. (2016). Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimers Dement.* 12, 292–323. doi: 10.1016/j.jalz.2016.02.002

Duff, K. (2014). One-week practice effects in older adults: tools for assessing cognitive change. *Clin. Neuropsychol.* 28, 714–725. doi: 10.1080/13854046.2014.920923

Duff, K., and Hammers, D. B. (2020). Practice effects in mild cognitive impairment: a validation of Calamia et al. (2012). *Clin. Neuropsychol.* [Epub Online ahead of print]. doi: 10.1080/13854046.2020.1781933

Duff, K., Foster, N. L., and Hoffman, J. M. (2014). Practice effects and amyloid deposition: preliminary data on a method for enriching samples in clinical trials. *Alzheimer Dis. Assoc. Disord.* 28:247. doi: 10.1097/WAD.0000000000000021

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., et al. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 19, 932–939. doi: 10.1097/JGP.0b013e318209dd3a

Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., and 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. *Alzheimers Dement.* 4, 11–18. doi: 10.1016/j.trci.2017.11.001

Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., et al. (2015a). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimers Dement.* 11, 415–424. doi: 10.1016/j.jalz.2014.03.005

Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Bondi, M. W., and Alzheimer's Disease Neuroimaging Initiative (2015b). Subtle Cognitive Decline and Biomarker Staging in Preclinical Alzheimer's Disease. *J. Alzheimers Dis. JAD* 47, 231–242. doi: 10.3233/JAD-150128

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., et al. (2018). Underdiagnosis of mild cognitive impairment: a consequence of ignoring practice effects. *Alzheimers Dement.* 10, 372–381. doi: 10.1016/j.dadm.2018.04.003

Elman, J. A., Panizzon, M. S., Gustavson, D. E., Franz, C. E., Sanderson-Cimino, M. E., Lyons, M. J., et al. (2020). Amyloid-β positivity predicts cognitive decline but cognition predicts progression to amyloid-β positivity. *Biol. Psychiatry* 87, 819–828. doi: 10.1016/j.biopsych.2019.12.021

Eppig, J. S., Edmonds, E. C., Campbell, L., Sanderson-Cimino, M., Delano-Wood, L., Bondi, M. W., et al. (2017). Statistically derived subtypes and associations with cerebrospinal fluid and genetic biomarkers in mild cognitive impairment: a latent profile analysis. *J. Int. Neuropsychol. Soc.* 23, 564–576. doi: 10.1017/S135561771700039X

Eppig, J., Werhane, M., Edmonds, E. C., Wood, L.-D., Bangen, K. J., Jak, A., et al. (2020). "Neuropsychological Contributions to the Diagnosis of Mild Cognitive Impairment Associated With Alzheimer's Disease. Vascular Disease," in *Alzheimer's Disease, and Mild Cognitive Impairment: advancing an Integrated Approach* eds D. J. Libon, M. Lamar, R. A. Swenson, and K. M. Heilman (Oxford: Oxford University Press), 52. doi: 10.1093/oso/9780190634230.003.0004

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement.* 1, 103–111. doi: 10.1016/j.dadm.2014.11.003

Gross, A. L., Anderson, L., and Chu, N. (2017). Do people with Alzheimer's disease improve with repeated testing? Unpacking the role of content and context in retest effects. *Alzheimers Dement.* 13, 473–474. doi: 10.1093/ageing/afy136

Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M. M., Sachs, B., et al. (2015). Predictors of retest effects in a longitudinal study of cognitive aging in a diverse community-based sample. *J. Int. Neuropsychol. Soc.* 21, 506–518. doi: 10.1017/S1355617715000508

Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J. Q., Bittner, T., et al. (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. *Alzheimers Dement.* 14, 1470–1481. doi: 10.1016/j.jalz.2018.01.010

Ho, D., Imai, K., King, G., Stuart, E., and Whitworth, A. (2018). *Package 'MatchIt'. Version 3*.

Jack, C. R. Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 14, 535–562. doi: 10.1016/j.jalz.2018.02.018

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

Jack, C. R. Jr., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., Knopman, D. S., et al. (2017). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimers Dement.* 13, 205–216. doi: 10.1016/j.jalz.2016.08.005

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., et al. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 17, 368–375. doi: 10.1097/JGP.0b013e31819431d5

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A., Jones, R. N., Choi, S. E., et al. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimers Dement.* 12:e12055. doi: 10.1002/dad2.12055

Malek-Ahmadi, M. (2016). Reversion from mild cognitive impairment to normal cognition. *Alzheimer Dis. Assoc. Disord.* 30, 324–330. doi: 10.1097/wad.0000000000000145

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

Mitchell, A. J., and Shiri-Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. *Acta Psychiatr. Scand.* 119, 252–265. doi: 10.1111/j.1600-0447.2008.01326.x

Olsson, B., Lautner, R., Andreasson, U., Öhrfelt, A., Portelius, E., Bjerke, M., et al. (2016). CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis. *Lancet Neurol.* 15, 673–684. doi: 10.1016/S1474-4422(16)00070-3

Pandya, S. Y., Clem, M. A., Silva, L. M., and Woon, F. L. (2016). Does mild cognitive impairment always lead to dementia? A review. *J. Neurol. Sci.* 369, 57–62. doi: 10.1016/j.jns.2016.07.055

Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., et al. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74, 201–209. doi: 10.1212/WNL.0b013e3181cb3e25

Rönnlund, M., and 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, 63–78. doi: 10.1016/j.intell.2005.06.004

Ronnlund, M., Nyberg, L., Backman, L., and 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, 3–18. doi: 10.1037/0882-7974.20.1.3

Salthouse, T. A. (2010). Selective review of cognitive aging. *J. Int. Neuropsychol. Soc.* 16, 754–760. doi: 10.1017/s1355617710000706

Sanderson-Cimino, M., Elman, J. A., Tu, X. M., Gross, A. L., Panizzon, M. S., Gustavson, D. E., et al. (2020). Cognitive Practice Effects Delay Diagnosis; Implications for Clinical Trials. *medRxiv* [Preprint]. doi: 10.1101/2020.11.03.20224808

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

Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., et al. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimers Res. Ther.* 3:32. doi: 10.1186/alzrt94

Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., et al. (2014). The A4 study: stopping AD before symptoms begin? *Sci. Transl. Med.* 6:228fs13. doi: 10.1126/scitranslmed.3007941

Sperling, R., Mormino, E., and Johnson, K. (2014). The evolution of preclinical Alzheimer's disease: implications for prevention trials. *Neuron* 84, 608–622. doi: 10.1016/j.neuron.2014.10.038

Thomas, K. R., Cook, S. E., Bondi, M. W., Unverzagt, F. W., Gross, A. L., Willis, S. L., et al. (2020). Application of neuropsychological criteria to classify mild cognitive impairment in the active study. *Neuropsychology* 34:862. doi: 10.1037/neu0000694

Thomas, K. R., Edmonds, E. C., Delano-Wood, L., and Bondi, M. W. (2017). Longitudinal trajectories of informant-reported daily functioning in empirically

defined subtypes of mild cognitive impairment. *J. Int. Neuropsychol. Soc.* 23, 521–527. doi: 10.1017/S1355617717000285

Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., et al. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement.* 15, 1322–1332. doi: 10.1016/j.jalz.2019.06.4948

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement.* 15, 106–152. doi: 10.1016/j.jalz.2018.08.005

Wang, G., Kennedy, R. E., Goldberg, T. E., Fowler, M. E., Cutter, G. R., and 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:e0228064. doi: 10.1371/journal.pone.0228064

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., et al. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. *J. Intern. Med.* 256, 240–246. doi: 10.1111/j.1365-2796.2004.01380.x

Conflict of Interest: MB receives royalties from Oxford University Press.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 03 June 2022
doi: 10.3389/fnagi.2022.885621

Parameterizing Practice in a Longitudinal Measurement Burst Design to Dissociate Retest Effects From Developmental Change: Implications for Aging Neuroscience

Nicholas Tamburri[1†], *Cynthia McDowell*[1,2†] *and Stuart W. S. MacDonald*[1,2*]

[1]*Brain Aging and Neurocognitive Health Laboratory, Department of Psychology, University of Victoria, Victoria, BC, Canada,* [2]*Institute on Aging and Lifelong Health, University of Victoria, Victoria, BC, Canada*

OPEN ACCESS

Edited by:
Daniel Nation,
University of California, Irvine,
United States

Reviewed by:
Denis Smirnov,
University of California, San Diego,
United States
Jonathan Hakun,
The Pennsylvania State University,
United States

***Correspondence:**
Stuart W. S. MacDonald
smacd@uvic.ca

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

Received: *28 February 2022*
Accepted: *05 May 2022*
Published: *03 June 2022*

Citation:
Tamburri N, McDowell C and MacDonald SWS (2022) Parameterizing Practice in a Longitudinal Measurement Burst Design to Dissociate Retest Effects From Developmental Change: Implications for Aging Neuroscience. Front. Aging Neurosci. 14:885621. doi: 10.3389/fnagi.2022.885621

Background: In longitudinal designs, the extraneous influence of retest effects can confound and obscure estimates of developmental change. The current study provides a novel approach to independently parameterize short-term retest effects and long-term developmental change estimates by leveraging a measurement burst design and three-level multilevel modeling. We further employ these short- and long-term slopes as predictors of cognitive status at long-term follow-up assessments.

Methods: Participants included 304 older adults from Project MIND: a longitudinal measurement burst study assessing cognitive performance across both biweekly sessions and annual retests. Participants were classified as either Healthy controls (HC) or Cognitively Impaired, not Demented (CIND) at baseline, the final burst assessment (Year 4), and at an additional four-year follow-up (Year 8). Response time inconsistencies (RTI) were computed at each burst occasion for a simple choice response time (CRT) task and a one-back response time (BRT) task. Three-level multilevel models were employed to simultaneously examine change in RTI for both CRT and BRT across weeks within years, as well as across years, in order to dissociate within-individual retest effects (short-term) from developmental (long-term) change slopes. Individual slopes were then extracted and utilized in a series of multinomial logistic regression equations to contrast short- vs. long-term RTI change as predictors of cognitive status.

Results: Separately parameterizing short- and long-term change estimates yielded distinct patterns of variation. CRT RTI remained stable across short-term weekly assessments, while significantly increasing across years. In contrast, BRT RTI decreased significantly across short-term assessments but showed no change across long-term assessments. After dissociating change estimates, short-term BRT as well as long-term CRT and BRT estimates predicted cognitive status at long-term follow-ups; increases in RTI, suggesting either an inability to benefit from retest or process-based developmental decline, were associated with an increased likelihood of being classified as CIND.

Conclusions: We showcase an innovative approach to dissociate retest effects from developmental change across and within individuals. Accurately parameterizing these distinct change estimates can both reduce systematic bias in longitudinal trend estimates as well as provide a clinically useful tool by utilizing retest effects to predict cognitive health and impairment.

Keywords: retest effects, practice vs. developmental change, longitudinal measurement burst design, cognitive aging, multilevel modeling (MLM)

PARAMETERIZING PRACTICE IN A LONGITUDINAL MEASUREMENT BURST DESIGN TO DISSOCIATE RETEST EFFECTS FROM DEVELOPMENTAL CHANGE: IMPLICATIONS FOR AGING NEUROSCIENCE

The analysis of change has posed numerous seemingly intractable problems for both clinicians and researchers studying human development, prompting contentions as to whether change could, or even should, be measured (e.g., Cronbach and Furby, 1970; Willett, 1988). Such debates motivated a fundamental conceptual shift in which developmental change became viewed as a continuous process that fluctuates over time, as opposed to mere increments between pre-post testing occasions (Willett, 1988). This reconceptualization, paired with Baltes and Nesselroade's (1979) assertion that one of the primary objectives of developmental research was to directly identify intraindividual change (i.e., exploring within-person processes), facilitated the development of increasingly sophisticated methodologies aimed at providing richer and more accurate parameterizations of between- and within-person change processes. The current study aims to further extrapolate upon these methodologies by employing innovative solutions to some of the more persistent problems inherent in modeling development.

Within aging neuroscience, where developmental outcomes are of central interest, longitudinal designs afford the opportunity to directly observe both age- and process-related change. Such designs allow researchers to avoid the biases inherent in cross-sectional inferences of change (see Baltes and Nesselroade, 1979; Hofer and Sliwinski, 2006; Schaie, 2008)—which employ between-subjects comparisons within age-heterogeneous samples to draw conclusions about the nature of age-graded development—and more appropriately approximate the conceptualization of change as a continuous and oscillatory process (Willett, 1988; Singer and Willett, 2003). However, while advances in conceptual and technical approaches have undoubtedly improved the ability to index change, many problems remain that continue to obfuscate the understanding and measurement of development.

Implicit in the reconceptualization of change as a continuous and intraindividual process is the understanding that change is modulated by a confluence of multiple influences occurring across both short and long temporal intervals. There is a pressing need to dissociate these processes, and their potentially confounding impact on true underlying development, to fully understand moderators of short- and long-term change. Of particular interest, retest effects—changes in performance attributable to previous exposure to the testing materials, environment, and procedures—perturb estimates of aging and development by systematically biasing inter- and intraindividual change trajectories in longitudinal designs (Hoffman et al., 2011). Retest effects, encompassing the more specific delineation of practice effects (i.e., improvements attributable to the repetition of the same or similar materials), are an oft-cited criticism of longitudinal designs and represent an enduring problem in the field of aging neuroscience (e.g., Schaie, 1965; Baltes, 1968). Retest effects have long been known to confound estimates of change across both short- (e.g., between first and second retest occasions) and long-term intervals (e.g., across many years of retest occasions; Thorndike et al., 1928; Ferrer et al., 2004; Wilson et al., 2006; Rabbitt et al., 2009). Given that longitudinal designs offer the only direct way of indexing intraindividual development, overcoming this susceptibility to retest effects is of critical importance to developmental researchers.

Appropriate quantification and parameterization of retest effects are crucial for understanding their unique value as an individual differences predictors. The magnitude of retest effects has shown to be differentiable depending on both test (e.g., complexity, modality) and test-taker characteristics (e.g., IQ, age, personality, mood, motivations; Bartels et al., 2010). The parameterization of retest effects may therefore serve as a useful cognitive variable, indicative of both an individual's current capacity and predicted cognitive trajectory. While the findings in this domain are equivocal, some evidence suggests that an individual's ability to benefit from practice is informative of their prospective cognitive health and disease risk—with smaller than expected practice effects in older adults potentially presaging cognitive decline, poorer response to intervention, and greater risk of Alzheimer's-related pathology (Duff et al., 2017; De Simone et al., 2021). For persons with mild cognitive impairment (MCI), inclusive of amnestic MCI (a-MCI), there is considerably more controversy as to whether these individuals can benefit from retest effects and, if so, across which cognitive domains (see Duff et al., 2017 for review). These contentions are further complicated as there is currently no widely accepted approach for reliably and accurately modeling variance due to retest. However, a recent investigation by De Simone et al. (2021) found that lacking the expected benefits from practice on episodic memory tests was an accurate prognostic indicator of late conversion to Alzheimer's disease in a-MCI patients. Distinguishing among

individuals who will remain stable a-MCI vs. progress to dementia is both a pressing objective and imposing challenge, given the known lability and heterogeneity of this relatively broad cognitive classification (Ganguli et al., 2004; Malek-Ahmadi, 2016). Among other benefits, innovations in parameterizing and dissociating retest from development could facilitate a deeper understanding of the utility of retest effects as sensitive predictors for distinguishing between- and within-person differences in cognitive function.

Formal attempts to control for retest effects have centered upon three basic approaches: (1) materials, (2) research design, and (3) quantitative modeling. A common method of material manipulation used by researchers—the use of alternate forms in cognitive testing—attempts to account for the most basic of practice effects (i.e., repeated exposure to the same testing material). However, this strategy has shown variable effectiveness depending on the construct being tested (Watson et al., 1994; Uchiyama et al., 1995; Benedict and Zgaljardic, 1998) and fails to address issues attributable to the more general impact of retest effects (e.g., previous exposure to the testing environment, procedure, etc.). Therefore, various longitudinal design considerations have been implemented to address this more encompassing definition of retest effects and control for their impact on developmental change.

Traditional longitudinal designs typically consist of widely-spaced measurement occasions (e.g., spanning years) in an effort to capture the timescale by which normative age-graded changes take place. However, in such instances, aging and retest effects are entirely conflated (e.g., 1-year increments in chronological age for a design specifying one-year retest intervals spanning five occasions), posing a particular challenge for modeling distinct and unbiased estimates of either process. Consequently, the failure to account for retest effects often leads to inaccurate characterizations of the rate and pattern of developmental change (e.g., change is underestimated), can cause violations of modeling assumptions (e.g., age convergence), and may undermine subsequent attempts of understanding change through regression or correlation analyses (Sliwinski et al., 2010a). More intricate longitudinal designs, such as waitlist control designs, attempt to address retest effects at the group level by employing a hold-out sample. Thorvaldsson et al. (2006), for example, utilized a waitlist control design to evaluate retest effects within several standardized cognitive performance domains. Initially, the researchers randomly selected one-third of their total sample to be assessed on their cognitive performance between the ages of 70–81. The remaining two-thirds of participants were prescribed as the hold out sample, to be assessed at a later date. From ages 85 to 99 years the cognitive performance of both the participants who were previously assessed (i.e., "original" participants), and a random selection of the remaining two-thirds of participants (i.e., "waitlist" participants), were then assessed concurrently. The comparison of cognitive performance between the original participants and waitlist participants facilitates an estimation of group-level retest effects. However, while this approach reasonably quantifies the average retest effects in a population, it precludes the investigation of intraindividual change and forces researchers to adopt questions of change that accommodate a between-person design (Thorvaldsson et al., 2006; Hoffman et al., 2011). Ultimately, when intraindividual change is of interest, controlling for retest *via* design decisions is exceptionally challenging. Indeed, the nature of repeated-measures data presumes the influence of retest effects as unavoidable (Salthouse's, 2013) and thus cannot be overcome by study design changes alone. Therefore, in addition to careful design considerations, adept statistical modeling approaches are also needed to more effectively address the impact of retest effects.

Advanced quantitative modeling techniques attempt to parse the effects of retest and aging into separately estimated model parameters. These quantitative approaches frequently consist of hierarchical or more sophisticated computational models (e.g., multilevel modeling, latent growth curve modeling, etc.) that estimate both maturational influences (e.g., aging) along with retest effects as separate parameters within a single analytic model of intraindividual change (e.g., Ferrer et al., 2004; Salthouse et al., 2004; Rabbitt et al., 2008). Although potentially informative, these modeling techniques remain subject to common, underappreciated pitfalls and assumptions that must be explicitly addressed. For instance, satisfying assumptions of age-convergence—that cross-sectional age differences and longitudinal age changes converge onto a common trajectory—is necessary in order to obtain meaningful parameter estimates of aging and retest. Hoffman et al. (2011) assert that failing to test and meet age-convergence assumptions can lead to significant bias and increased Type 1 error rates in the estimation of retest effects. This is particularly the case for traditional longitudinal designs that often leverage equal interval designs where age and retest occasion are perfectly correlated. Disconcertingly, most studies that attempt to directly model retest effects often fail to explicitly test for age-convergence assumptions (Sliwinski et al., 2010b). Furthermore, while retest models attempt to estimate a "test naïve" aging trajectory that is dissociated from retest effects, these models are, in actuality, estimating aging trajectories by holding retest effects constant across time. This implicit assumption, that retest effects are invariant in magnitude across time, is potentially spurious when considering that retest effects are (1) often most pronounced between the first and second measurement occasion (Collie et al., 2003; Bartels et al., 2010; Scharfen et al., 2018), (2) potentially affected by ceiling effects (Calamia et al., 2012), (3) influenced by individual differences in the amount and rate of time-dependent forgetting (MacDonald et al., 2006), and (4) showcase interindividual differences in magnitude dependent on test- and test-taker variables (Bartels et al., 2010). Thus, while the combination of both analytical and methodological advances has clearly informed the extant literature, there are notable gaps remaining vis-a-vis optimal approaches for effectively distinguishing retest effects from change.

Researchers are evidently presented with numerous permutations of both design and analytic strategies that provide differential advantages and disadvantages when investigating longitudinal change in cognition; however, when dissociating and parameterizing retest effects is of critical interest, a

recent recommendation suggests combining the advantages of the seldom-used measurement burst design alongside the well-known utility of multilevel modeling (Sliwinski, 2008; Hoffman et al., 2011; Jones, 2015). Measurement burst designs can explicitly measure retest effects by examining variability across both short-term intervals—such as narrowly spaced retests (e.g., daily, weekly) in which meaningful age-based change is unlikely to occur—as well as long-term periods (e.g., yearly) over which durable age-based developmental changes commonly unfold. This design avoids common pitfalls of more traditional longitudinal designs, including concerns of age-convergence and equal-interval measurement occasions, and provides the opportunity for more nuanced statistical analysis. Specifically, multilevel modeling can be used to partial these distinct levels of variability into separate slope parameters, separately estimating and dissociating the impact of short-term retest-related change from more durable developmental change.

Unfortunately, many current investigations of retest effects employ two-level multilevel models for a research objective that is optimally addressed using three-level nested data. Specifically, for measurement burst designs and multilevel modeling to be utilized effectively for modeling retest, the innovative application of three-level multilevel models is required to systematically dissociate variance within-persons across short-term retest occasions (level 1) and long-term developmental intervals (level 2), as well as between-persons (level 3). Investigating two-level models by inappropriately aggregating three-level data not only yields an inaccurate dissociation of retest and developmental change but also generates criticism regarding the leveraging of short-term retest intervals as proxies for retest effects altogether. Salthouse's (2013), for example, has suggested that employing short-term slopes as indices of retest effects is contingent upon having identified positive, moderately strong associations between short- and long-term change estimates—an intuitive assumption given the expectation that shorter-term retest gains should be positively linked to longer-term developmental increases as well. In contrast to this expectation, Salthouse's (2013) reported a modest negative association between retest and long-term change in cognition. Notably, however, these findings were based upon a two-level analysis of change (i.e., a latent change analysis) from a data set characterized by at least three nested levels—sessions (level 1), within occasions (level 2), within persons (level 3). Failing to properly account for the nestedness inherent within a dataset can result in parameter estimates that are confounded with extraneous sources of information and violate modeling assumptions (e.g., data dependency) which can result in inaccurate probability estimates and confounded estimates of short- and long-term change. This is especially problematic when the research questions and/or conclusions are predicated upon having accurately quantified variance at select levels. Thus, when considering the viability of using short-term change slopes as indicators of retest effects, utilizing a measurement burst design and a three-level modeling framework will provide a more accurate dissociation and quantification of retest and developmental variance.

Using data from Project Mental Inconsistency in Normals and Dementia (MIND), an innovative longitudinal measurement burst design study, the current study employed advanced quantitative models to dissociate short-term retest effects and long-term developmental change and investigated the relative predictivity of retest and change for differentiating cognitive status subgroups at long-term follow-up assessments. Given that retest and developmental change represent non-independent time structures, we utilized three-level multilevel modeling to separately estimate within-individual change in cognitive function across short-term weekly retests (level 1) and long-term yearly bursts (level 2), as well as between-individual differences (level 3) in cognitive performance. The use of a three-level hierarchical modeling structure, paired with the previously suggested measurement burst design, represents a critical extension of the existing literature that simultaneously parameterizes within-person change across both short-term biweekly assessments (i.e., retest) as well as across longer-term annual assessments (i.e., developmental age-based change). Specific research objectives included: (1) disaggregating short- (weekly) from long-term (annual) change slopes to estimate and empirically evaluate the patterns and association among these estimates of retest and development; and (2) leveraging these dissociable estimates of change, obtained during the course of the 4-year measurement burst study, as independent individual-differences predictors of cognitive status indexed at Year 4 (the conclusion of the burst design) and Year 8 (the conclusion of the Project MIND study). The first objective was accomplished by investigating change in response time inconsistencies (RTI) for two select cognitive measures—a simple choice response time (CRT) task and a more complex 1-back choice response time (BRT) task—using three-level multilevel models. By specifying random effects in these multilevel models, it was possible to derive person-specific change slopes that were extracted to address our second research question which used multinomial logistic regression models to contrast short- and long-term RTI change as predictors of cognitive status at Year 4 and 8 of the study.

Increasing evidence suggests that RTI represents a dissociable dimension of performance relative to mean Response Time (RT) (MacDonald and Stawski, 2015, 2020) that may better capture underlying changes in physiological and cognitive processes (Dixon et al., 2007; de Ribaupierre and Lecerf, 2018). Previous research also suggests that within-person variability is differentially sensitive to cognitive status groups, such that RTI was most pronounced in subjects with more severe cognitive impairment (Strauss et al., 2007; MacDonald and Stawski, 2020). The utilization of RTI is particularly beneficial for the current investigation that leverages lability in cognitive performance—which is particularly sensitive to retest effects and generally resistant to floor and ceiling effects—as a proxy for cognitive health status.

MATERIALS AND METHODS

Participants

Participants were 304 community-dwelling Caucasian older adults aged 64–92 years (M = 74.02; SD = 5.95) who were concerned about their cognitive functioning but had not been

diagnosed with a neurological disorder. This study was approved by the University of Victoria Human Research Ethics Board and was conducted in accordance with institutional guidelines. Participants (208 female and 96 male) resided in Victoria, Canada and were recruited through local media advertisements (radio and newspaper). Participants were generally well-educated (M = 15.15; SD = 3.14; range = 7–24 years of education), performed well on the Mini-Mental State Examination (MMSE; Folstein et al., 1975) (M = 28.74; SD = 1.23; range = 24–30), and were in relatively good health (total number of chronic health conditions: M = 2.92; SD = 1.91; range = 0–10). Exclusionary criteria at intake included physician-diagnosed dementia or an MMSE score of less than 24, drug or alcohol abuse, psychotropic drug use, current psychiatric diagnosis, a history of significant head injury (e.g., loss of consciousness greater than 5 min), other neurological or major medical illnesses (e.g., Parkinson's disease, cancer, heart disease), severe sensory impairment (e.g., difficulty reading newspaper-size print, difficulty hearing a normal conversation), and lack of fluency in English.

Procedure

Participants were initially screened for inclusion and exclusion criteria *via* a telephone interview. Baseline testing occurred across seven sessions (one group and six individuals) scheduled over approximately 3 months. The group testing session was held at the university in our laboratories and the individual testing sessions were conducted in the participant's home. The first two sessions (one group and one individual) were used to obtain demographic and health information and to administer cognitive measures. Participants then completed a burst evaluation, consisting of five individual biweekly testing sessions that varied across days of the week and times of the day. Within these sessions, participants completed various assessments including cognitive performance measures such as RT tasks that were designed to assess short-term fluctuations in response speed. The entire testing battery was repeated annually four times. During each annual wave, the cognitive measures (inclusive of the burst RT tasks) were identical, and the order of presentation did not vary. However, for each subsequent year after baseline, four (rather than five) biweekly testing sessions were completed, yielding up to 17 total assessments for each individual (see **Figure 1**). Follow-up demographic and cognitive assessments were then conducted four years following cessation of the burst portion of the study (i.e., at Year 8) to evaluate long-term change in participants' cognitive status. Eighty percent of participants (N = 242) completed all four bursts and attrition rates were 11.0%, 3.5%, and 4.5% of the sample between years 1–2, 2–3, and 3–4, respectively. The attrition rate between Year 4 and Year 8 was 26%, with 61% of the original sample (N = 185) completing Year 8.

Cognitive Status

Cognitive status was ascertained for each year of study according to participant's performance on five cognitive tasks. The cognitive performance tasks consisted of indicators for perceptual speed (WAIS-R Digit Symbol Substitution; Wechsler, 1981), verbal fluency (Controlled Associations; Ekstrom et al., 1976), vocabulary (Extended Range Vocabulary; Ekstrom et al., 1976), episodic memory (Immediate free word recall; Hultsch et al., 1990), and inductive reasoning (Letter Series; Thurstone, 1962). Participants were classified as cognitively intact healthy controls (HC) or cognitively impaired, not demented (CIND) based upon deficits (1.5 SDs relative to age and education norms) spanning the five distinct cognitive domains. The age and education norms were obtained from 445 adults aged 65–94 years from the Victoria Longitudinal Study (Dixon and de Frias, 2004); this normative comparison sample for deriving cognitive status classifications was partitioned into four age and education groups (age = 65–74 years and 75+ years; education = 0–12 years and 13+ years) with means and standard deviations computed for each of the five cognitive reference measures. Participants classified as CIND were further subdivided as CIND-S based on deficits for a single cognitive measure or as CIND-M based on deficits across two or more of the cognitive reference tasks. A more thorough methodological account of Project MIND, inclusive of the testing and cognitive status classification procedures, can be found in Bielak et al. (2010).

Response Time

RT tasks were presented on a Panasonic CF-48 laptop computer (Intel Pentium III 800-MHz processor, MS-DOS operating system Version 4.10.2222) with a 14″ color screen. The computer processor controlled the stimulus presentation and timing for each RT task. Participants responded to stimuli by pressing keys on a custom-designed response console consisting of an aluminum enclosure encompassing four response keys in a linear array. This response box was interfaced with the laptop through a PCMCIA Game Port, directly accessible by the CPU, in order to ensure millisecond timing latency (±1 ms). The RT tasks were programmed using C++ and were run on MS-DOS.

Choice Response Time (CRT)

Participants were presented with four plus signs displayed in a horizontal row along with a response input device containing four spatially-mapped keys. On each trial, following a 1,000 ms delay, a box replaced one of the plus signs. For each trial, participants were asked to respond to the location of the box as quickly as possible. Ten practice trials were followed by 60 test trials. The response latencies of the 60 test trials were used for analysis (Bielak et al., 2010).

One-Back Choice Response Time (BRT)

The BRT task used the same display, response box, and stimulus presentation design as the CRT task. However, for each trial, participants were asked to respond to the location of the box on the "previous" trial. A total of 10 practice trials followed by 61 test trials were administered. As participants did not respond on Trial 1, it was omitted and only the response latencies of the remaining 60 test trials were used for analysis.

Data Preparation

Outliers and Missing Values

All RT data were examined for outliers by examining the distributions of raw latency scores at the individual level.

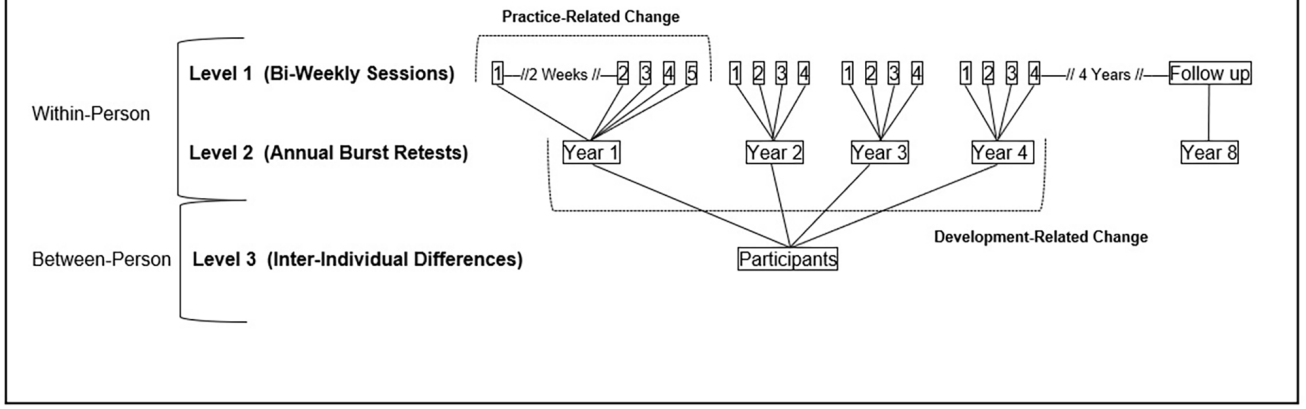


FIGURE 1 | Parameterization of retest and developmental change using three-level multilevel modeling and a measurement burst design.

Exceptionally slow or fast responses were removed and considered likely to represent sources of measurement error (e.g., accidental key press). Valid lower bound response times have been provided by previous research (150 ms; Hultsch et al., 2002), and valid upper bounds were identified by calculating intraindividual means and standard deviations for each task and measurement occasion; for each individual, any trials that exceeded their personal mean by three or more standard deviations were removed. For each of the CRT and BRT tasks at Year 1, a total of 91,200 trials were possible across individual assessments (60 trials per administration of each RT task), sessions (five biweekly retests), and persons (n = 304; 60 × 5 × 304 = 91,200). For the CRT task, 0.13% of trials were excluded due to missing values, 1.43% due to incorrect responses, and 1.78% due to trimming outliers, leaving 96.65% usable trials. For the BRT task, 0.20% of trials were excluded due to missing values, 10.46% due to incorrect responses, and 2.42% due to trimming outliers, leaving 86.93% usable trials. This data preparation procedure for eliminating outliers represents a conservative approach to examining intraindividual variability in RT performance by reducing within-subject variation.

Computation of Response Time Inconsistency (RTI)

RTI was indexed using residualized intraindividual standard deviation (ISD) estimates. The residualized ISD estimates were computed across RT trials for each session and burst, residualizing select confounds from the raw data by fitting a multilevel model in order to dissociate within- and between-subject sources of variation (MacDonald and Stawski, 2020). Removing systematic confounds yields RTI estimates that are not conflated with mean age differences in response speed, developmental change, or practice effects at the trial-to-trial level (Stawski et al., 2019; MacDonald and Stawski, 2020). For each session and burst, the computed residualized ISD scores were then linearly transformed into standardized T scores (M = 50, SD = 10). See Hultsch et al. (2008) for a full description of this procedure.

Statistical Procedure

The nested three-level data structure for the present study is characterized by weekly assessments (level 1) nested within annual bursts (level 2) nested within persons (level 3). Using the "nlme" package in R (Pinheiro et al., 2022), we addressed the first research objective by fitting three-level multilevel models to predict change in RTI for both CRT and BRT across sessions (biweekly assessments), bursts (annual retests), and persons. Multilevel modeling decomposes total variability into within- vs. between-person sources. Moreover, this multilevel framework, coupled with the current measurement burst design, facilitates parsing of intraindividual variability from intraindividual change (Nesselroade, 2002), thereby separately yet simultaneously indexing retest effects and developmental change, respectively.

Variance decomposition in CRT and BRT RTI across weeks, years, and persons was based upon preliminary fully unconditioned models. Two independent, conditioned longitudinal models were then fit to examine change in CRT and BRT RTI separately. Equation 1 demonstrates the modeling of average linear change in CRT RTI as a function of weekly and yearly assessments (fixed slope effects) and the variability of change across individuals (random slope effects). Response time inconsistencies on the CRT task (CRT RTI_{ijk}), for a given week (i), year (j), and person (k), were modeled as a function of that individual's performance at baseline testing, plus their average individual rate of change per each additional week and year examined (the slopes), plus an error term (ε). A number of random effects were also modeled, with the level-1 residuals [$Var(\varepsilon_{ijk})$] reflecting within-person week-to-week variability, and the level-2 residuals [$Var(\mu_{0jk})$] indexing within-person variability across the annual retest bursts. Variance in the level-3 residuals [$Var(\upsilon_{00k})$] index between-person stable variability averaged across all biweekly retests and annual burst assessments.

Select fixed effects of interest include population estimates for the average CRT RTI score (γ_{000}), the average biweekly retest (i.e., practice) effect (γ_{100}) as well as the average yearly retest (i.e., long-term developmental change) effect (γ_{010}).

Equation 1

$$
\begin{aligned}
\textit{Level 1} \quad CRT\,RTI_{ijk} &= \beta_{0jk} + \beta 1_{jk} Week_{ijk} + \varepsilon_{ijk} \\
\textit{Level 2} \quad \beta_{0jk} &= \delta_{00k} + \delta_{01k} Year_{jk} + \mu_{0jk} \\
\beta_{1jk} &= \delta_{10k} + \mu_{1jk} \\
\textit{Level 3} \quad \delta_{00k} &= \gamma_{000} + \gamma_{001} Age_k + \gamma_{002} Sex_k + \upsilon_{00k} \\
\delta_{01k} &= \gamma_{010} + \gamma_{011} Age_k + \gamma_{012} Sex_k + \upsilon_{01k} \\
\delta_{10k} &= \gamma_{100} + \upsilon_{01k}
\end{aligned}
$$

Weekly (Level 1) and yearly (Level 2) linear effects were centered at baseline (e.g., the first week for Year 1). Person (Level 3) covariates included age at baseline (γ_{001}; centered at age 74) and sex (γ_{002}; centered as 0 = males/1 = females) Parameter estimates were derived using full information maximum likelihood (FIML) estimation, using all available data under the assumption of missing at random (MAR; Grand et al., 2016).

For the second research question, we employed polytomous (multinomial) logistic regression to examine changes in RTI (both short-and long-term) as predictors of cognitive status in Year 4 and Year 8. Healthy controls served as the referent group for each model. Due to the small values of bi-weekly change (i.e., retest) estimates for both CRT and BRT (a consequence of millisecond temporal scaling), we rescaled these values as seconds to facilitate the interpretability of model point estimates and odds ratios.

RESULTS

Patterns of Retest and Development Across Time

Sample characteristics are reported in **Table 1**. To address the first research objective, and to provide an index of the data dependency inherent in our repeated measures design, unconditional models were first fit to decompose the total variability into within-person (weekly and yearly) and between-person sources. Of the total variability in CRT RTI across the sample, 68% reflected variability between-persons, whereas 6% and 26% reflected within-person variability across years and weeks, respectively. Comparable patterns were found for BRT RTI in which 75% of the total variability was between-persons, 7% within-persons across years, and 18% within-persons across weeks.

Conditioned longitudinal change analyses were then fit using three-level multilevel models to dissociate retest (i.e., short-term) from developmental (i.e., long-term) change estimates. Specifically, these models derived separate estimates of within-person change across both weeks and years, with the former estimate indexing change due to retest and the latter change due to developmental processes (see **Figure 1**). Between-person differences at baseline and across years were also explored, with random intercept and slope effects estimated to facilitate the derivation of individual slopes for use in subsequent logistic regression equations.

TABLE 1 | Sample characteristics as a function of baseline cognitive status.

Baseline	HC *N* = 136	CIND-S *N* = 88	CIND-M *N* = 80
Age (years)	73.3 (5.4)	73.8 (6.0)	75.5 (6.6)
Sex (% males)	29	26	43
Education (years)	15.9 (3.1)	15.2 (3.1)	14.3 (3.2)
MMSE score	29.0 (1.0)	28.7 (1.1)	28.3 (1.5)
[a]Medications	5.8 (3.5)	5.4 (3.3)	6.5 (8.9)
[b]Risk Factor (% without)	84	83	73
CIND Classification Year 4	*N* = 138	*N* = 62	*N* = 45
CIND Classification Year 8	*N* = 112	*N* = 40	*N* = 33

[a]*Self-reported number of total prescribed medications.* [b]*Presence of risk factor (Significant Hearing Loss, Neurological and/or Cardiac Condition). HC, Healthy Controls; CIND-S, Cognitively Impaired, not Demented based on single task deficit; CIND-M, Cognitively Impaired, not Demented based on >2 task deficits; MMSE, Mini-Mental State Examination (Folstein et al., 1975).*

Two separate models, controlling for age and sex, were fit to evaluate CRT RTI and BRT RTI independently as cognitive outcomes. Analyses revealed notable differences between retest and developmental change parameter estimates within each model, as well as between the two models. Specifically, population estimates for the CRT outcome model indicated non-significant short-term change in RTI (β = -0.02, $p > 0.05$), with this stability across week-to-week assessments connoting the absence of practice effects. However, RTI significantly increased across years in the study (β = 0.15, $p = 0.005$), demonstrating increasing cognitive variability in CRT performance over longer developmental trajectories. In contrast, our BRT model yielded an inverse pattern, perhaps reflecting the inherent differences in cognitive demands between the BRT and CRT measures. Within BRT, a task that requires higher-order cognitive processes (e.g., executive functioning), significant short-term declines in RTI (β = -0.06, $p < 0.0001$) exemplified the expected benefits of practice in reducing performance inconsistencies across week-to-week assessments. Yet, non-significant change in long-term RTI slopes (β = −0.09, $p > 0.05$) demonstrated stability in patterns of BRT consistency across years. Regardless of RT task, short and long-term change slopes—reflecting the presence of retest vs. developmental change—yielded distinct sources of information. Of note, age significantly predicted between-person differences in both the CRT (β = 0.21, $p < 0.001$) and BRT (β = 0.28, $p < 0.001$) tasks, such that increasing age resulted in increased inconsistencies for each RT task. Sex did not significantly predict RTI ($p > 0.05$) in either model. The fixed effects from these conditioned models are displayed in **Table 2**.

Retest and Developmental Change as Predictors of Cognitive Status

The aforementioned results summarized our fixed effects which describe the aggregated rates of change in the sample. However, we also estimated random effects in order to derive individual estimates of short- and long-term change for use as predictors of cognitive status. Specifically, to assess whether individual differences in retest effects and developmental change predicted cognitive status at long-term follow ups (Years 4 and 8), person-specific residuals were used to derive individual slope

TABLE 2 | Fixed effects for CRT and BRT three-level multilevel models.

	CRT RTI			BRT RTI		
Predictors	β	CI	*p*	β	CI	*P*
Intercept	7.44	6.91–7.96	<0.001	7.23	6.57–7.89	<0.001
Short-Term	−0.02	−0.05–0.00	0.082	−0.06	−0.09 to −0.04	<0.001
Long-Term	0.15	0.05–0.25	0.005	−0.09	−0.02–0.02	0.124
Age	0.21	0.16–0.26	<0.001	0.34	0.28–0.40	<0.001
Sex	−0.24	−0.87–0.38	0.439	0.25	−0.54–1.04	0.534

estimates for entry as predictors in several multinomial logistic regression models. These models investigated whether individual differences in short- and long-term rates of change in CRT and BRT RTI were predictive of CIND status upon conclusion of the burst portion of Project MIND (Year 4), as well as at the termination of the study (Year 8). At both Year 4 and Year 8 follow-up assessments, four separate multinomial logistic regression models, controlling for age and sex, were fit for each of our four RTI-related predictors: short-term CRT RTI, short-term BRT RTI, long-term CRT RTI, and long-term BRT RTI. These models were fit independently to avoid potential issues of collinearity between the short- and long-term slope estimates within each cognitive measure. Parameter estimates for these RTI predictors are presented in **Table 3**.

Short-term change slopes in CRT, indexing retest effects in the present study, were not significantly predictive of CIND status at either Year 4 or Year 8. However, short-term practice-related gains in BRT RTI were significantly associated with an increased likelihood of being classified as CIND-S [OR = 2.26, 95% CI (1.31, 3.88), $p = 0.003$] and CIND-M [OR = 3.82, 95% CI (2.14, 6.84), $p < 0.001$] at Year 4, as well as CIND-M at Year 8 [OR = 2.50, 95% CI (1.26, 4.98), $p = 0.009$].

In contrast, elevated yearly RTI was associated with increased odds of being classified as CIND relative to HC for both CRT and BRT. Long-term developmental slope estimates for CRT RTI were significantly associated with increased odds of being classified as CIND-M [OR = 4.33, 95% CI (1.68, 11.05), $p = 0.002$] at Year 4, with no significant associations at Year 8. Thus, holding constant age and sex differences, year-to-year unit increases in CRT RTI increased the likelihood of being classified as CIND-M over healthy controls by 333 percent. Additionally, unit increases in yearly BRT RTI were associated with an increased likelihood of being classified as CIND-S [OR = 2.05, 95% CI (1.16, 3.62), $p = 0.014$] and CIND-M [OR = 3.10, 95% CI (1.70, 5.68), $p < 0.001$] at Year 4, as well as CIND-M at Year 8 [OR = 2.23, 95% CI (1.04, 4.77), $p = 0.039$].

To further inform these patterns, four separate multinomial logistic regression models were fit using person-level baseline MMSE scores to contrast the predictivity of long-term cognitive status with our residualized RTI slope parameters. Specifically, we were interested in identifying whether a simple baseline cognitive measure would significantly contribute to model fit or show comparatively accurate long-term predictions of cognitive health status. Across all models, baseline MMSE performance neither significantly contributed to model fit nor predicted cognitive status at long-term follow ups, underscoring the utility of retest effects as more sensitive prognostic indices of cognitive health.

Our models also identified age as a significant predictor of CIND status, with increasing age generally facilitating an increased likelihood of being classified as cognitively impaired. Specifically, at Year 4, age significantly predicted both CIND-S and CIND-M for three of four models (with the exception of yearly CRT RTI which predicted CIND-M only). At Year 8, age was a significant predictor of CIND-M only, regardless of RT task or weekly or yearly RTI. Depending on the model, older age significantly predicted cognitive status such that each additional year beyond age 74 resulted in a 5%–10% increased likelihood of cognitive impairment, relative to controls. Sex (male or female) did not significantly predict cognitive status in any of the eight models.

Finally, to further delineate associations between individual slopes of short- and long-term change, we computed simple bivariate correlations for both CRT and BRT RTI. Correlations between short- and long-term individual BRT RTI slopes were significant and strong at the two-tailed level ($r = 0.87$, $p < 0.001$). For CRT RTI, short- and long-term change slopes shared a more modest but still significant association ($r = 0.39$, $p < 0.001$).

DISCUSSION

The current investigation showcases an innovative approach for studying practice effects in community-dwelling older adults using both novel design considerations and advanced statistical methodology. By utilizing a measurement burst design—in which data were collected across weeks within years, as well as across years—and employing three-level multilevel models, we were able to (a) dissociate short-term retest effects from long-term developmental change, (b) demonstrate that within-person change across these varying temporal intervals yields distinct patterns of variation, and (c) leverage these retest and change slopes as predictors of cognitive impairment status. The difference in slope estimates between short- and long-term change, and their respective predictive utility, highlights both (a) the advantage of the current approach for dissociating retest effects from developmental change, as well as (b) the promise of employing retest as a proxy for individual differences in cognitive health.

Retest Can Bias Estimates of Developmental Change

An enduring criticism of longitudinal research concerns the presence of retest effects which may obfuscate the magnitude, shape, and even estimated direction of developmental change. Although retest-related gains are thought to bias development and age-related changes in cognitive performance (Wilson et al.,

TABLE 3 | Multinomial logistic regression: weekly and annual RTI in relation to the likelihood of cognitive impairment status at Year 4 and Year 8.

	Year 4						Year 8					
	CIND-S			CIND-M			CIND-S			CIND-M		
Variable	OR	95% CI LB	95% CI UB	OR	95% CI LB	95% CI UB	OR	95% CI LB	95% CI UB	OR	95% CI LB	95% CI UB
Weekly CRT RTI	1.36	0.80	2.34	1.16	0.62	2.17	1.78	0.93	3.40	1.30	0.61	2.77
Yearly CRT RTI	1.81	0.76	4.34	4.33*	1.70	11.05	1.55	0.48	4.98	0.93	0.27	3.22
Weekly BRT RTI	2.26*	1.31	3.88	3.82**	2.14	6.84	1.94	0.99	3.81	2.50*	1.26	4.98
Yearly BRT RTI	2.05*	1.16	3.62	3.10**	1.70	5.68	1.68	0.81	3.51	2.28*	1.04	4.77

*Note. Age is baseline age centered at 74 years. Sex is categorically coded with females (1) as the reference category. Healthy controls (HC) serve as the reference category. CIND-S, cognitively impaired-not demented for a single cognitive outcome; CIND-M, cognitively impaired-not-demented for two or more cognitive outcomes; LB, lower bound; UB, upper bound; CRT, choice reaction time; BRT, 1-back choice reaction time; RTI, response time inconsistency. *$p < 0.05$; **$p < 0.001$.*

2006; Hoffman et al., 2011; MacDonald and Stawski, 2020), retest effects are seldom systematically measured or controlled for, due in part to the limitations of existing designs and quantitative methodologies (Sliwinski and Mogle, 2008; Salthouse's, 2013). Therefore, novel longitudinal approaches that consider the impact of retest effects and the utilization of advanced modeling approaches are needed to adequately distinguish within-person developmental change from retest-related change.

We investigated the extent to which weekly change (i.e., influenced by retest effects) and yearly change (i.e., influenced by aging and development) reflect comparable or distinct sources of information. Consistent with expectations, distinct and divergent patterns were present between the weekly short-term and annual long-term change slopes in both RT tasks. Non-significant, stable change in RTI in the CRT task over short retest intervals was differentiated from significant long-term performance declines. This is consistent with the understanding that simple psychomotor abilities (e.g., sensorimotor speed, processing speed) are less susceptible to the influence of retest and showcase normative declines with aging (Salthouse, 1996; Duff et al., 2017). For RTI in the BRT task, our sample showed the expected benefits of retest with significant short-term performance gains but demonstrated non-significant change over longer retest intervals. These patterns are also congruent with previous research, as the BRT task—which draws upon more executive processes (e.g., updating)—is increasingly susceptible to practice-related gains pursuant to repeated exposure (Grand et al., 2016). The use of such a task helps bolster the idea that placing more demands on cognitive processing resources may provide a more sensitive evaluation of retest effects. Such disparate patterns observed in the fixed effects for both the simpler CRT task, and the more cognitively demanding BRT task, indicate that the within-person change slopes across weekly and annual temporal intervals reflect non-redundant sources of information. Neglecting to parse cognitive performance according to these distinct time structures would bias slope estimates, confounding retest effects with developmental change. These results corroborate previous research demonstrating the important and considerable impact of retest on developmental change slopes (e.g., Wilson et al., 2006; Hoffman et al., 2011; Jones, 2015) and suggest that related but unique associations exist between these constructs. Moreover, overlooking the potential influence of retest effects may mask underlying cognitive symptomatology or early detection of cognitive decline.

The non-significant developmental slope in BRT RTI may, despite our systematic parsing of short-term retest-related variance from long-term parameter estimates, be indicative of the more enduring, generalized impact of retest—which has shown to exert influence across much longer retest (e.g., years) intervals (Rabbitt et al., 2004; Salthouse et al., 2004). However, an alternative explanation is that the observed long-term BRT RTI stability is a consequence of collapsing individual performance information across all cognitive status groups onto one linear trajectory. The heterogeneity in cognitive status produces diverging trajectories of RTI among CIND subgroups (see MacDonald and Stawski, 2020), yet yields a relatively flat sample average slope when combined. Notably, the shape and magnitude of the sample average slope are less consequential to our key research focus, which is concerned with evaluating whether a) there are individual differences in short- and long-term change, and b) these individual differences in slopes are linked to cognitive status. Therefore, the choice to model the data as an average slope, irrespective of cognitive status (i.e., not including a CIND status moderator), was intentional in order to derive person-specific slopes (reflecting individual deviations in change from the population average) which could predict cognitive status at long-term follow-ups.

Utilizing Retest as a Predictor of Prospective Cognitive Impairment

The focus of our second research objective was to investigate whether the unique intraindividual slopes derived from our models were predictive of cognitive health outcomes (i.e., CIND), for as many as four years following the completion of cognitive testing. We examined within-person change directly by investigating whether an individual's short-term retest slope predicted long-term cognitive status, and whether their developmental slope reflected a reliable index of process-based change (dissociable from short-term change). A series of multinomial logistic regression models were used to contrast short- vs. long-term CRT and BRT RTI as individual predictors of cognitive status at the final burst assessment wave (Year 4) as well as an additional four years later at the conclusion of the study (Year 8). These follow-up assessments correspond with the natural middle and endpoints of the study and afford a novel opportunity to investigate the differential predictive utility of our discrete cognitive slopes.

Using this approach, we demonstrated that the likelihood of being classified as CIND-M relative to HC at Year 4 was over three times greater for individuals showcasing annual increases in CRT RTI. This result, including the non-significant predictive ability of short-term CRT RTI change, is consistent with the extant literature on psychomotor function and decline. Specifically, the basic sensorimotor demands of the comparatively less cognitively demanding CRT task resulted in less intraindividual variability and diminished retest from which to accurately predict long-term cognitive status. However, interindividual differences in annualized intraindividual change may be reflective of unique intraindividual processes (e.g., normative or pathological aging) or characteristics (e.g., health-related comorbidities) that facilitate more accurate predictions of cognitive impairment status at long-term follow-up (Stawski et al., 2015). Although annualized CRT RTI was not significantly predictive of cognitive status at Year 8, this may be due in part to the relative heterogeneity and lability of CIND classifications or the relative insensitivity of developmental CRT RTI as a proxy for underlying bio-cognitive dysfunction.

In contrast to CRT RTI, increases in intraindividual BRT RTI across both weeks and years were significantly predictive of cognitive status at Years 4 and 8. These patterns reflect the expected influence of both retest and developmental performance on long-term cognitive status. Individuals who failed to benefit from retest and exhibited increases in their short-term BRT RTI were significantly more likely to be classified as CIND-S or CIND-M at Year 4, as well as CIND-M at Year 8. These predictive patterns support the potential clinical utility of retest, where the ability to benefit from practice is postulated to be a function of underlying cognitive health (Galvin et al., 2005; Duff et al., 2011, 2012). Long-term increases in annual intraindividual BRT RTI were also associated with increased odds of being classified as CIND-S and CIND-M at Year 4, and CIND-M at Year 8. Independent of age and sex differences, individuals characterized by increasing BRT RTI across short- and long-term intervals were associated with an increased likelihood of cognitive impairment classification. The identical trends between weekly and annual increases in BRT RTI underscore a key finding of our study: when appropriately parameterized, both intraindividual retest and developmental change slopes can yield distinguishable and meaningful predictions of long-term health outcomes.

Retest as an Early Indicator of Cognitive Decline

The observed discrepancies between the CRT and BRT tasks are consistent with previous research indicating that retest effects are test-specific (Benedict and Zgaljardic, 1998; Wilson et al., 2006). In comparison to the CRT task, the BRT task involves increased cognitive demands that likely involve attention, working memory, and inhibitory control which are more sensitive to retest effects (Grand et al., 2016). This dependency on executive processes not only underscores why BRT RTI is more sensitive to retests effects but may also help elucidate why both short- and long-term BRT performance showcased greater predictive accuracy for classifying CIND status at Year 8.

More generally, RTI holds considerable promise as a sensitive marker of normal and pathological aging and has received much attention for its promise as a proxy for central nervous system (CNS) health and an early indicator of cognitive impairment or decline (Hultsch et al., 2000; Bielak et al., 2010; MacDonald et al., 2011; MacDonald and Stawski, 2020). RTI has been shown to predict late-life deleterious health outcomes (e.g., fall risk, vascular impairment, dementia; for review, see MacDonald and Stawski, 2015) and may enhance our understanding of the dynamic relationship between individual fluctuations in cognitive performance and underlying CNS integrity (Halliday et al., 2017). RTI has also garnered empirical support as an indicator of lapses of attention (particularly for tasks requiring executive control processes; West et al., 2002), processing efficiency (Eysenck and Calvo, 1992; Brose et al., 2010), and has been shown to fluctuate depending on perceived competence in cognitive control (e.g., individual differences in control beliefs for age-related changes in cognitive performance). For example, in a recent investigation of RTI in both CRT and BRT measures, Cerino et al. (2020) identified that increases in perceived competence were associated with lower RTI on the CRT task, and higher (i.e., maladaptive) RTI performance on the BRT task in older adults. Taken together, BRT RTI may serve as a unique cognitive health indicator, sensitive to disruptions in executive function attributable to both labile (e.g., momentary fluctuations in attention reflecting mental noise, daily variations in sleep or distress) or more chronic mechanisms (e.g., pathological aging, dopaminergic dysregulation, declining CNS signaling fidelity) affecting higher-order cognition. In the context of the present study, increased RTI for the BRT (vs. CRT) task may be a more effective proxy for these underlying bio-cognitive disturbances, which may account for BRT's increased predictivity at both the level of retest and development across longer time periods.

Accordingly, RTI across both short-and long-term follow-up intervals demonstrated stronger predictivity for differentiating CIND-M from HC, compared to CIND-S; this dose-response pattern was expected given that the CIND-M classification represents deficits across multiple cognitive domains and was more likely to include impairments in executive function. Findings from the logistic regression models also speak to the known lability of cognitive impairment classifications. Specifically, the clinical trajectory of CIND is frequently recognized as unstable and heterogeneous, with several studies demonstrating that single-domain cognitive impairment classifications (e.g., CIND-S) are associated with higher instability and increased likelihood of reverting to HC compared to multi-domain classifications (e.g., Diniz et al., 2009; Loewenstein et al., 2009; Ritchie and Tuokko, 2010; Sachdev et al., 2013).

Associations Between Short- and Long-Term Change

The ability for retest effects to be leveraged as early indicators of cognitive decline is predicated upon having accurately parameterized retest-effect-related variance, as well as the idea that short-term retest occasions are associated with long-term developmental trajectories. As some theorists have argued, short intervals may only serve as estimates of retest effects in longitudinal designs when the associations between short- and long-term change are positive and at least moderately

strong. For example, Salthouse's (2013) reported how short- and long-term changes across several cognitive domains were negatively correlated, whereby individuals showing the greatest short-term gains (between first and second sessions within one occasion) also exhibited the largest longer-term losses (across occasions). On the basis of such negative associations, it has been questioned whether short-term slopes can be reliably used as estimates of retest effects (or correspondingly as individual-differences predictors) in longitudinal models. Notably, however, Salthouse's (2013) criticism was based upon a two-level latent change analysis (assessments within persons) of three-level nested data (sessions within occasions within persons)—a fact that raises concerns about the impact of between-context dependency on the direction, magnitude, and significance of the reported short- and long-term change estimates (estimated separately as two-level structures as opposed to derived simultaneously in three-level models) as well as their negative association.

To circumvent these concerns, in the present study we employed a novel three-level approach that more accurately parameterized short- and long-term change estimates, prior to deriving unbiased estimates of the association between retest and developmental change. For CRT RTI, we found a significant positive correlation between short- and long-retest intervals. This moderate correlation is indicative of shifts in the rank-order association between changes in short- vs. long-term CRT RTI change estimates, with the lack of significant short-term retest effects presaging the non-significant predictivity of Year 4 and 8 cognitive status. For BRT RTI, we identified a large-magnitude positive correlation between short- and long-term intervals; those who exhibited greater increases in variability across short-term retests (i.e., benefitted less from practice) also exhibited greater annualized developmental increases in RTI (a known indicator of various deleterious, age-related outcomes; MacDonald and Stawski, 2015). The increased association shown in BRT RTI further supports the potential utility of modeling short-term intervals as retest effects in longitudinal models, and is consistent with the reported susceptibility of BRT to retest-related effects (Bielak et al., 2010; Grand et al., 2016). This correlation also corroborates our logistic regression results, where individuals who benefited more from practice were also more likely to be cognitively intact at long-term follow-up. For both CRT and BRT RTI, the association between retest and developmental change slopes was positive and robust. These results are in keeping with the findings of other researchers who have advocated for the utility of short-term intervals as a proxy for retest effects and identified robust positive correlations between short- and long-term change (Zimprich et al., 2004; Hoffman et al., 2011).

Implications for Aging Neuroscience

Our results highlight several notable implications for research on cognitive aging and the cognitive neuroscience of aging: (1) increases in RTI, even on simple psychomotor tasks, are associated with an increased risk of cognitive impairment up to four to eight years post-baseline assessment; (2) long-term developmental trajectories in cognition, while not substantially different from short-term trajectories, yield larger odds of being subsequently classified as cognitively impaired; and (3) individuals who not only fail to benefit from expected retest-related gains but also worsen in performance across years are at increased odds of being classified as cognitively impaired at follow-up. This latter result is consistent with previous literature asserting that retest effects can be a useful indicator of cognitive decline (Duff et al., 2011, 2012; Jutten et al., 2020). In the present study, the predictive utility of short- and long-term slope estimates to independently discriminate among cognitive status groups, even as many as four years later, speaks to the promise of individual differences in change for distinct time structures as predictors of future cognitive impairment. By combining modern design and analytics, researchers can systematically disaggregate short- from long-term within-person variability and utilize unbiased estimates of retest and developmental change to predict cognitive health and impairment. By using retest effects as a proxy for cognitive health, practitioners and individuals may be able to track inconsistencies across short-term temporal intervals, reducing the need for rigorous annual cognitive neuropsychological testing batteries. Harnessing the predictive validity of retest effects, by accurately parsing them from developmental effects, can serve as a clinically useful, non-invasive, and inexpensive tool for earlier detection and increasing diagnostic accuracy of cognitive impairment. Appropriate forethought and parameterization of retest effects are therefore paramount to both reduce systematic bias in longitudinal trend estimates as well as harness the unique opportunity that retest effects offer as individual-differences predictors.

Study Strengths and Limitations

The current study showcases several strengths including the exploration of differing psychomotor tasks based on lower and higher-order cognitive demands (i.e., CRT vs. BRT), sufficient sample sizes for each cognitive status classification, the 8-year duration of the study permitting the examination of cognitive impairment status for both near and distal follow-up periods, as well as the use of performance variability (i.e., RTI) which has been suggested to serve as an important proxy of CNS integrity (Halliday et al., 2017). The present findings replicate previous research on the clinical utility of RTI (e.g., MacDonald and Stawski, 2020), as well as the predictive utility of retest effects over shorter intervals (e.g., Duff et al., 2012; Jutten et al., 2020) as early markers of shifts in cognitive health. Furthermore, previous researchers have suggested the use of multilevel modeling and intensive repeated measures burst designs for addressing retest effects in developmental research (e.g., Nesselroade, 1991; Sliwinski, 2008; Salthouse and Nesselroade, 2010; Sliwinski et al., 2010a; Salthouse's, 2013); this study is among the first to combine such intricate design recommendations along with appropriately matching quantitative analyses (three-level multilevel modeling) for deriving unbiased estimates of retest and their corresponding prediction of cognitive status.

To be sure, this study is not without limitations. First, cognitive status was determined using a battery of neuropsychological measures and a distributional CIND classification, rather than by clinical interview. Additionally, cognitive status classifications

were determined by performance (below 1.5 SDs based on age and education-matched peers) on the number of tasks; classifications were not determined by the nature of cognitive impairment (i.e., amnestic vs. non-amnestic) and therefore this study could not address etiology-specific impairment subtypes (MacDonald and Stawski, 2020). Second, the sample was fairly homogeneous and composed of relatively healthy, well-educated individuals which may restrict generalizability. Notwithstanding, we were able to distinguish between cognitive subgroups in this sample which highlights the robustness of our findings. It is likely that a more heterogeneous, less healthy sample would produce even stronger results. It is also recommended that additional research employ this design and modeling approach to prospectively identify whether individual differences in retest slopes can predict cognitive impairment or dementia progression, without *a priori* knowledge of cognitive groupings. Finally, our model's long-term developmental change estimates may remain biased by retest, given that mere-exposure effects have been shown to exert influence even across longer retest intervals spanning years (Rabbitt et al., 2004; Salthouse et al., 2004). However, the significant predictive ability of our individual developmental slope estimates for BRT RTI in identifying individuals at risk of being CIND-S and CIND-M at long-term follow-ups highlights the utility of these slopes as predictors of cognitive status, irrespective of whether corresponding long-term increases in RTI are slightly underestimated due to generalized practice effects that span much longer retest intervals.

Future Directions

Investigators seeking to further explore dissociable patterns between retest and development should consider modeling non-linear trends across short- and long-term trajectories. Additionally, exploring whether retest effects can significantly predict subtypes of CIND (e.g., non-amnestic vs. amnestic CIND) will further elucidate the utility of retest effects as sensitive indicators of cognitive decline. Finally, whereas the present study focused on RTI, future investigations may utilize our approach to explore the dissociable patterns of retest and developmental change using other common metrics (e.g., central tendency, accuracy) for cognitive function.

CONCLUSIONS

The present study overviews an innovative approach for parameterizing retest effects in longitudinal designs where developmental outcomes in older adulthood are of interest. We leveraged an intensive repeated measurement burst design as well as three-level multilevel modeling to operationalize retest and developmental change directly and distinctly in the same model. Such an approach generates more definitive, less confounded trajectories of change by disaggregating within-person short- and long-term cognitive performance estimates. Further, when investigating the predictive utility of short- and long-term change in cognitive variability, we demonstrated that both retest effects and developmental change estimate each independently predicted cognitive status, thereby highlighting their potential clinical utility as well as underscoring the importance of accurately parameterizing both retest and developmental change in longitudinal designs. Specifically, for measures implicating executive functioning (i.e., BRT), individuals who fail to benefit from the expected influence of retest and instead exhibit both short- and long-term increases in RTI are at an increased risk of being classified as cognitively impaired up to 4 years post data collection. Researchers and clinicians alike may adopt the synergistic advantages of the measurement burst design and three-level multilevel modeling to facilitate better parameterization of retest and developmental effects and improved predictivity of cognitive function. In doing so, retest effects may serve as a clinically useful tool for predicting prospective cognitive status without the need for overly long or intensive neuropsychological testing batteries.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: participants of this study did not agree for their data to be shared publicly, so supporting data is not available. Requests to access these datasets should be directed to smacd@uvic.ca.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Victoria Human Research Ethics Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NT, CM, and SM all contributed to the conceptual design and implementation of the research. SM reviewed and edited the manuscript and supervised the study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (06468-2017) to SM. Original data collection for Project MIND was supported by grants from the Canadian Institutes of Health Research (CIHR) to David F. Hultsch, the Alzheimer Society of Canada to Esther Strauss, and the Natural Sciences and Engineering Research Council of Canada to David F. Hultsch and Esther Strauss.

ACKNOWLEDGMENTS

SM also graciously acknowledges support of the Royal Society of Canada's College of New Scholars, Artists and Scientists. Further information about Project MIND can be obtained from SM (smacd@uvic.ca).

REFERENCES

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Hum. Dev.* 11, 145–171. doi: 10.1159/000270604

Baltes, P. B., and Nesselroade, J. R. (1979). "History and rationale of longitudinal research," in *Longitudinal Research in the Study of Behavior and Development*, eds J. R. Nesselroade and P. B. Baltes (New York, NY: Academic Press), 1–39.

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., and Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neurosci.* 11:118. doi: 10.1186/1471-2202-11-118

Benedict, R. H., and Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *J. Clin. Exp. Neuropsychol.* 20, 339–352. doi: 10.1076/jcen.20.3.339.822

Bielak, A. A. M., Hultsch, D. F., Strauss, E., MacDonald, S. W. S., and Hunter, M. A. (2010). Intraindividual variability in reaction time predicts cognitive outcomes 5 years later. *Neuropsychology* 24, 731–741. doi: 10.1037/a0019802

Brose, A., Schmiedek, F., Lövdén, M., Molenaar, P. C., and Lindenberger, U. (2010). Adult age differences in covariation of motivation and working memory performance: contrasting between-person and within-person findings. *Res. Hum. Dev.* 7, 61–78. doi: 10.1080/15427600903578177

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Cerino, E. S., Stawski, R. S., Geldhof, G. J., and MacDonald, S. (2020). Associations between control beliefs and response time inconsistency in older adults vary as a function of attentional task demands. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 75, 1819–1830. doi: 10.1093/geronb/gby124

Collie, A., Maruff, P., Darby, D. G., and McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *J. Int. Neuropsychol. Soc.* 9, 419–428. doi: 10.1017/S1355617703930074

Cronbach, L. J., and Furby, L. (1970). How we should measure "change": or should we? *Psychol. Bull.* 74, 68–80. doi: 10.1037/h0029382

de Ribaupierre, A., and Lecerf, T. (2018). On the importance of intraindividual variability in cognitive development. *J. Intell.* 6:17. doi: 10.3390/jintelligence6020017

De Simone, M. S., Perri, R., Rodini, M., Fadda, L., De Tollis, M., Caltagirone, C., et al. (2021). A lack of practice effects on memory tasks predicts conversion to Alzheimer disease in patients with amnestic mild cognitive impairment. *J. Geriatr. Psychiatry Neurol.* 34, 582–593. doi: 10.1177/0891988720944244

Diniz, B. S., Nunes, P. V., Yassuda, M. S., and Forlenza, O. V. (2009). Diagnosis of mild cognitive impairment revisited after one year. Preliminary results of a prospective study. *Dement. Geriatr. Cogn. Disord.* 27, 224–231. doi: 10.1159/000203346

Dixon, R. A., and de Frias, C. M. (2004). The Victoria longitudinal study: from characterizing cognitive aging to illustrating changes in memory compensation. *Aging Neuropsychol. Cogn.* 11, 346–376. doi: 10.1080/13825580490511161

Dixon, R. A., Garrett, D. D., Lentz, T. L., MacDonald, S. W. S., Strauss, E., and Hultsch, D. F. (2007). Neurocognitive resources in cognitive impairment: exploring markers of speed and inconsistency. *Neuropsychology* 21, 381–399. doi: 10.1037/0894-4105.21.3.381

Duff, K., Atkinson, T. J., Suhrie, K. R., Dalley, B. C., Schaefer, S. Y., and Hammers, D. B. (2017). Short-term practice effects in mild cognitive impairment: evaluating different methods of change. *J. Clin. Exp. Neuropsychol.* 39, 396–407. doi: 10.1080/13803395.2016.1230596

Duff, K., Callister, C., Dennett, K., and Tometich, D. (2012). Practice effects: a unique cognitive variable. *Clin. Neuropsychol.* 26, 1117–1127. doi: 10.1080/13854046.2012.722685

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., et al. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 19, 932–939. doi: 10.1097/JGP.0b013e318209dd3a

Ekstrom, R. B., French, J. W., Harman, H. H., and Dermen, D. (1976). *Manual for Kit of Factor Referenced Cognitive Tests.* Princeton, NJ: Educational Testing Service.

Eysenck, M. W., and Calvo, M. G. (1992). Anxiety and performance: the processing efficiency theory. *Cogn. Emot.* 6, 409–434. doi: 10.1080/02699939208409696

Ferrer, E., Salthouse, T. A., Stewart, W. F., and Schwartz, B. S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychol. Aging* 19, 243–259. doi: 10.1037/0882-7974.19.2.243

Folstein, M., Folstein, S., and McHugh, S. (1975). Mini-mental state: a practical method for grading the cognitive status of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Galvin, J. E., Powlishta, K. K., Wilkins, K., McKeel, D. W., Jr., Xiong, C., Grant, E., et al. (2005). Predictors of preclinical Alzheimer disease and dementia: a clinicopathologic study. *Arch. Neurol.* 62, 758–765. doi: 10.1001/archneur.62.5.758

Ganguli, M., Dodge, H. H., Shen, C., and DeKosky, S. T. (2004). Mild cognitive impairment, amnestic type: an epidemiologic study. *Neurology* 63, 115–121. doi: 10.1212/01.wnl.0000132523.27540.81

Grand, J. H., Stawski, R. S., and MacDonald, S. W. (2016). Comparing individual differences in inconsistency and plasticity as predictors of cognitive function in older adults. *J. Clin. Exp. Neuropsychol.* 38, 534–550. doi: 10.1080/13803395.2015.1136598

Halliday, D. W., Stawski, R. S., and MacDonald, S. W. (2017). Cognitively-impaired-not demented status moderates the time-varying association between finger tapping inconsistency and executive performance. *Arch. Clin. Neuropsychol.* 32, 110–116. doi: 10.1093/arclin/acw084

Hofer, S. M., and Sliwinski, M. J. (2006). "Design and analysis of longitudinal studies on aging," in *Handbook of the Psychology of Aging*, eds J. E. Birren and K. W. Schaire (San Diego, CA: Elsevier Inc./Academic Press), 15–37. doi: 10.1016/B978-012101264-9/50005-7

Hoffman, L., Hofer, S. M., and Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: a simulation study. *Psychol. Aging* 26, 778–791. doi: 10.1037/a0023910

Hultsch, D. F., Hertzog, C., and Dixon, R. A. (1990). Ability correlates of memory performance in adulthood and aging. *Psychol. Aging* 5, 356–368. doi: 10.1037//0882-7974.5.3.356

Hultsch, D. F., MacDonald, S. W. S., and Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 57, 101–115. doi: 10.1093/geronb/57.2.p101

Hultsch, D. F., MacDonald, S. W., Hunter, M. A., Levy-Bencheton, J., and Strauss, E. (2000). Intraindividual variability in cognitive performance in older adults: comparison of adults with mild dementia, adults with arthritis and healthy adults. *Neuropsychology* 14, 588–598. doi: 10.1037//0894-4105.14.4.588

Hultsch, D. F., Strauss, E., Hunter, M. A., and MacDonald, S. W. S. (2008). "Intraindividual variability, cognition and aging," in *The Handbook of Aging and Cognition*, eds F. I. M. Craik and T. A. Salthouse (New York, NY: Psychology Press), 3, 491–556.

Jones, R. N. (2015). Practice and retest effects in longitudinal studies of cognitive functioning. *Alzheimers Dement. (Amst)* 1, 101–102. doi: 10.1016/j.dadm.2015.02.002

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S., Jones, R. N., Choi, S. E., et al. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimers Dement. (Amst)* 12:e12055. doi: 10.1002/dad2.12055

Loewenstein, D. A., Acevedo, A., Small, B. J., Agron, J., Crocco, E., and Duara, R. (2009). Stability of different subtypes of mild cognitive impairment among the elderly over a 2- to 3-year follow-up period. *Dement. Geriatr. Cogn. Disord.* 27, 418–423. doi: 10.1159/000211803

MacDonald, S. W. S., DeCarlo, C. A., and Dixon, R. A. (2011). Linking biological and cognitive aging: toward improving characterizations of developmental time. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 66, i59–i70. doi: 10.1093/geronb/gbr039

MacDonald, S. W. S., and Stawski, R. S. (2015). "Intraindividual variability—An indicator of vulnerability or resilience in adult development and aging?," in *The Handbook of Intraindividual Variability Across the Life Span*, eds M. Diehl, K. Hooker and M. Sliwinski (New York, NY: Routledge), 231–257.

MacDonald, S. W. S., and Stawski, R. S. (2020). Longitudinal changes in response time mean and inconsistency exhibit predictive dissociations for risk of cognitive impairment. *Neuropsychology* 34, 264–275. doi: 10.1037/neu0000608

MacDonald, S. W., Stigsdotter-Neely, A., Derwinger, A., and Bäckman, L. (2006). Rate of acquisition, adult age and basic cognitive abilities predict forgetting: new views on a classic problem. *J. Exp. Psychol. Gen.* 135, 368–390. doi: 10.1037/0096-3445.135.3.368

Malek-Ahmadi, M. (2016). Reversion from mild cognitive impairment to normal cognition: a meta-analysis. *Alzheimer Dis. Assoc. Disord.* 30, 324–330. doi: 10.1097/WAD.0000000000000145

Nesselroade, J. R. (1991). "The warp and woof of the developmental fabric," in *Visions of Aesthetics, the Environment and Development: The Legacy of Joachim F. Wohwill*, eds R. Downs, L. Liben and D. S. Palermo (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 213–240.

Nesselroade, J. R. (2002). Elaborating the different in differential psychology. *Multivariate Behav. Res.* 37, 543–561. doi: 10.1207/S15327906MBR3704_06

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2022). *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-155.* Available online at: https://CRAN.R-project.org/package=nlme.

Rabbitt, P., Diggle, P., Holland, F., and McInnes, L. (2004). Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 59, P84–P97. doi: 10.1093/geronb/59.2.p84

Rabbitt, P., Lunn, M., Ibrahim, S., and McInnes, L. (2009). Further analyses of the effects of practice, dropout, sex, socio-economic advantage and recruitment cohort differences during the University of Manchester longitudinal study of cognitive change in old age. *Q. J. Exp. Psychol. (Hove)* 62, 1859–1872. doi: 10.1080/17470210802633461

Rabbitt, P., Lunn, M., and Wong, D. (2008). Death, dropout and longitudinal measurements of cognitive change in old age. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 63, 271–278. doi: 10.1093/geronb/63.5.p271

Ritchie, L. J., and Tuokko, H. (2010). Patterns of cognitive decline, conversion rates and predictive validity for 3 models of MCI. *Am. J. Alzheimers Dis. Other Demen.* 25, 592–603. doi: 10.1177/1533317510382286

Sachdev, P. S., Lipnicki, D. M., Crawford, J., Reppermund, S., Kochan, N. A., Trollor, J. N., et al. (2013). Factors predicting reversion from mild cognitive impairment to normal cognitive functioning: a population-based study. *PLoS One* 8:e59649. doi: 10.1371/journal.pone.0059649

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychol. Rev.* 103, 403–428. doi: 10.1037/0033-295x.103.3.403

Salthouse's, T. A. (2013). Effects of age and ability on components of cognitive change. *Intelligence* 41, 501–511. doi: 10.1016/j.intell.2013.07.005

Salthouse, T. A., and Nesselroade, J. R. (2010). Dealing with short-term fluctuation in longitudinal research. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 65, 698–705. doi: 10.1093/geronb/gbq060

Salthouse, T. A., Schroeder, D. H., and Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Dev. Psychol.* 40, 813–822. doi: 10.1037/0012-1649.40.5.813

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychol. Bull.* 64, 92–107. doi: 10.1037/h0022371

Schaie, K. W. (2008). "Historical processes and patterns of cognitive aging," in *Handbook of Cognitive Aging: Interdisciplinary Perspectives*, eds S. M. Hofer and D. F. Alwin (Thousand Oaks, CA: Sage Publications, Inc.), 368–383. doi: 10.4135/9781412976589.n23

Scharfen, J., Peters, J. M., and Holling, H. (2018). Retest effects in cognitive ability tests: a meta-analysis. *Intelligence* 67, 44–66. doi: 10.1016/j.intell.2018.01.003

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195152968.001.0001

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Soc. Personal. Psychol. Compass* 2, 245–261. doi: 10.1111/j.1751-9004.2007.00043.x

Sliwinski, M., Hoffman, L., and Hofer, S. (2010a). "Modeling retest and aging effects in a measurement burst design," in *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development*, eds P. C. M. Molenaar and K. M. Newell (Washington, DC: American Psychological Association), 37–50. doi: 10.1037/12140-003

Sliwinski, M., Hoffman, L., and Hofer, S. M. (2010b). Evaluating convergence of within-person change and between-person age differences in age-heterogeneous longitudinal studies. *Res. Hum. Dev.* 7, 45–60. doi: 10.1080/15427600903578169

Sliwinski, M. J., and Mogle, J. (2008). "Time-based and process-based approaches to analysis of longitudinal data," in *Handbook of Cognitive Aging: Interdisciplinary Perspectives*, eds S. M. Hofer and D. F. Alwin (Thousand Oaks, CA: Sage Publications, Inc.), 477–491. doi: 10.4135/9781412976589.n28

Stawski, R. S., MacDonald, S., Brewster, P., Munoz, E., Cerino, E. S., and Halliday, D. (2019). A comprehensive comparison of quantifications of intraindividual variability in response times: a measurement burst approach. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 74, 397–408. doi: 10.1093/geronb/gbx115

Stawski, R. S., Smith, J., and MacDonald, S. W. S. (2015). "Intraindividual variability and covariation across domains in adulthood and aging: contributions for understanding behavior, health, and development," in *Handbook of Intraindividual Variability Across the Life Span*, eds M. Diehl, K. Hooker and M. J. Sliwinski (New York, NY: Routledge/Taylor and Francis Group), 258–279.

Strauss, E., Bielak, A. A., Bunce, D., Hunter, M. A., and Hultsch, D. F. (2007). Within-person variability in response speed as an indicator of cognitive impairment in older adults. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* 14, 608–630. doi: 10.1080/13825580600932419

Thorndike, E. L., Bregman, E. O., Tilton, J., and Woodyard, E. (1928). *Adult Learning.* New York, NY: Macmillan.

Thorvaldsson, V., Hofer, S. M., Berg, S., and Johansson, B. (2006). Effects of repeated testing in a longitudinal age-homogeneous study of cognitive aging. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 61, P348–P354. doi: 10.1093/geronb/61.6.p348

Thurstone, T. G. (1962). *Primary Mental Abilities: Grades 9-12, 1962 Revision.* Chicago, IL: Science Research Associates.

Uchiyama, C. L., D'Elia, L. F., Dellinger, A. M., Becker, J. T., Selnes, O. A., Wesch, J. E., et al. (1995). Alternate forms of the Auditory-Verbal Learning Test: issues of test comparability, longitudinal reliability and moderating demographic variables. *Arch. Clin. Neuropsychol.* 10, 133–145. doi: 10.1016/0887-6177(94)E0034-M

Watson, F. L., Pasteur, M.-A. L., Healy, D. T., and Hughes, E. A. (1994). Nine parallel versions of four memory tests: an assessment of form equivalence and the effects of practice on performance. *Hum. Psychopharmacol. Clin. Exp.* 9, 51–61. doi: 10.1002/hup.470090107

Wechsler, O. (1981). *Wechsler Adult Intelligence Scale - Revised, Manual.* New York, NY: Psychological Corporation.

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I., and Stuss, D. T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain Cogn.* 49, 402–419. doi: 10.1006/brcg.2001.1507

Willett, J. B. (1988). Chapter 9: questions and answers in the measurement of change. *Rev. Res. Educ.* 15, 345–422. doi: 10.3102/0091732X015001345

Wilson, R. S., Li, Y., Bienias, J. L., and Bennett, D. A. (2006). Cognitive decline in old age: separating retest effects from the effects of growing older. *Psychol. Aging* 21, 774–789. doi: 10.1037/0882-7974.21.4.774

Zimprich, D., Hofer, S. M., and Aartsen, M. J. (2004). Short-term versus long-term longitudinal changes in processing speed. *Gerontology* 50, 17–21. doi: 10.1159/000074384

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 16 June 2022
doi: 10.3389/fnagi.2022.883131

Avoid or Embrace? Practice Effects in Alzheimer's Disease Prevention Trials

Andrew J. Aschenbrenner[1], Jason Hassenstab[1], Guoqiao Wang[1], Yan Li[1], Chengjie Xiong[1], Eric McDade[1], David B. Clifford[1], Stephen Salloway[2], Martin Farlow[3], Roy Yaari[4], Eden Y. J. Cheng[4], Karen C. Holdridge[4], Catherine J. Mummery[5], Colin L. Masters[6], Ging-Yuek Hsiung[7], Ghulam Surti[8], Gregory S. Day[9], Sandra Weintraub[10], Lawrence S. Honig[11], James E. Galvin[12], John M. Ringman[13], William S. Brooks[14], Nick C. Fox[15], Peter J. Snyder[8], Kazushi Suzuki[16], Hiroyuki Shimada[17], Susanne Gräber[18] and Randall J. Bateman[1] for the Dominantly Inherited Alzheimer Network Trials Unit (DIAN-TU)*

[1] Washington University in St. Louis School of Medicine, St. Louis, MO, United States, [2] Warren Alpert Medical School of Brown University, Providence, RI, United States, [3] Indiana University School of Medicine, Indianapolis, IN, United States, [4] Eli Lilly and Company, Indianapolis, IN, United States, [5] University College London, London, United Kingdom, [6] University of Melbourne, Melbourne, VIC, Australia, [7] The University of British Columbia, Vancouver, BC, Canada, [8] The University of Rhode Island, Kingston, RI, United States, [9] Mayo Clinic, Jacksonville, FL, United States, [10] Feinberg School of Medicine, Northwestern University, Chicago, IL, United States, [11] Columbia University Irving Medical Center, New York, NY, United States, [12] Miller School of Medicine, University of Miami, Miami, FL, United States, [13] University of Southern California, Los Angeles, CA, United States, [14] Neuroscience Research Australia, University of New South Wales Medicine, Randwick, NSW, Australia, [15] Dementia Research Center, University College London, London, United Kingdom, [16] The University of Tokyo, Tokyo, Japan, [17] Osaka City University, Osaka, Japan, [18] German Center for Neurodegenerative Disease (DZNE), Tübingen, Germany

OPEN ACCESS

Edited by:
Daniel Nation, University of California, Irvine, United States

Reviewed by:
David Morgan, Michigan State University, United States
Sheila Black, The University of Alabama, United States
Daniel Gillen, University of California, Irvine, United States

***Correspondence:**
Andrew J. Aschenbrenner
a.aschenbrenner@wustl.edu

Specialty section:
This article was submitted to Alzheimer's Disease and Related Dementias, a section of the journal Frontiers in Aging Neuroscience

Received: *24 February 2022*
Accepted: *19 May 2022*
Published: *16 June 2022*

Citation:
Aschenbrenner AJ, Hassenstab J, Wang G, Li Y, Xiong C, McDade E, Clifford DB, Salloway S, Farlow M, Yaari R, Cheng EYJ, Holdridge KC, Mummery CJ, Masters CL, Hsiung G-Y, Surti G, Day GS, Weintraub S, Honig LS, Galvin JE, Ringman JM, Brooks WS, Fox NC, Snyder PJ, Suzuki K, Shimada H, Gräber S and Bateman RJ (2022) Avoid or Embrace? Practice Effects in Alzheimer's Disease Prevention Trials. Front. Aging Neurosci. 14:883131. doi: 10.3389/fnagi.2022.883131

Demonstrating a slowing in the rate of cognitive decline is a common outcome measure in clinical trials in Alzheimer's disease (AD). Selection of cognitive endpoints typically includes modeling candidate outcome measures in the many, richly phenotyped observational cohort studies available. An important part of choosing cognitive endpoints is a consideration of improvements in performance due to repeated cognitive testing (termed "practice effects"). As primary and secondary AD prevention trials are comprised predominantly of cognitively unimpaired participants, practice effects may be substantial and may have considerable impact on detecting cognitive change. The extent to which practice effects in AD prevention trials are similar to those from observational studies and how these potential differences impact trials is unknown. In the current study, we analyzed data from the recently completed DIAN-TU-001 clinical trial (TU) and the associated DIAN-Observational (OBS) study. Results indicated that asymptomatic mutation carriers in the TU exhibited persistent practice effects on several key outcomes spanning the entire trial duration. Critically, these practice related improvements were larger on certain tests in the TU relative to matched participants from the OBS study. Our results suggest that the magnitude of practice effects may not be captured by modeling potential endpoints in observational studies where assessments are typically less frequent and drug expectancy effects are absent. Using alternate instrument forms (represented in our study by computerized tasks) may partly mitigate

practice effects in clinical trials but incorporating practice effects as outcomes may also be viable. Thus, investigators must carefully consider practice effects (either by minimizing them or modeling them directly) when designing cognitive endpoint AD prevention trials by utilizing trial data with similar assessment frequencies.

Keywords: practice effects, Alzheimer's disease, clinical trials, learning, assessment frequency, alternative forms

INTRODUCTION

Phase 3 secondary prevention clinical trials in Alzheimer's disease aim to demonstrate the efficacy of drug or other interventions in preserving or improving cognitive function in at-risk individuals. Such trials typically use the slowing of the rate of cognitive decline between a treatment arm and a placebo group as their primary efficacy endpoint (Sperling et al., 2014; Bateman et al., 2017; Cummings et al., 2020). Comprehensive neuropsychological test batteries are administered at regular intervals (e.g., every 6–12 months) to best characterize cognitive change across the course of the trial and to monitor for adverse events such as unexpected drops in performance. However, these repeated administrations may have unanticipated consequences for trial outcomes. Specifically, it is well-known that healthy adults typically improve in performance (termed "practice effects" or "PEs") with repeated cognitive testing (Calamia et al., 2012). These PEs can be attributed to several factors including increased familiarity with task procedures, development of testing strategies, or memorization of specific stimuli. These gains are not limited to short time intervals and can persist for as long as 7 years (Salthouse et al., 2004) after just one exposure, a longer time span than a typical AD prevention trial. It is also important to consider that in symptomatic AD populations, where active neurodegenerative processes drive worsening cognitive performance, practice effects do not always translate to better performance from visit to visit. Rather, the competing forces of disease and PEs can manifest as attenuations of decline such that PEs may be observable as flat or simply less negative slopes.

For these reasons, potential PEs must be taken into consideration when planning a clinical trial. The two primary analytical models used in AD trials either analyze change from baseline to final test (e.g., mixed models for repeated measures or MMRM) or conceptualize change as linear from baseline to end of study (random intercept and slope models). When PEs are present but unaccounted for in statistical models, the magnitude of decline over the course of the trial can be drastically underestimated (Hassenstab et al., 2015; Jacobs et al., 2017) reducing the power to detect a treatment effect. Therefore, it may be desirable to minimize the influence of PEs in a clinical trial. One way to do so would be to include multiple "screening" sessions (Goldberg et al., 2015) which give participants experience with the cognitive battery prior to the initiation of treatment, as PEs tend to be largest after the first or second retest (Collie et al., 2003; Bartels et al., 2010). Other methods for minimizing PEs include the use of alternate forms, although this presents the additional challenge of verifying that the different forms are truly psychometrically equivalent (Gross et al., 2012), and yet still limit PEs due to familiarity. Computerized cognitive assessments, depending on the test paradigms, can protect against form-related PEs by randomly selecting stimuli for each test administration, creating an essentially endless number of alternate forms. But of course, this requires additional equipment and study management that can be costly and may not suit all trial protocols. Importantly, none of these approaches are entirely successful at eliminating practice effects (Beglinger et al., 2005; Falleti et al., 2006). Given the difficulties with eliminating PEs in cognitive studies, some studies have turned away from efforts at avoiding PEs opting instead to determine if incorporating PEs as outcomes themselves may reveal meaningful information about cognitive status. For example, several studies have shown that the attenuation of PEs in clinically healthy older adults can predict important outcomes such as biomarker status or risk of progression to symptomatic AD (Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2017; Oltra-Cucarella et al., 2018; Samaroo et al., 2020). PEs may therefore serve as a subtle marker of early disease even if average cognitive trajectories are relatively flat. It is critical, therefore, to have a comprehensive understanding of factors that produce or exaggerate practice effects and to develop statistical tools to appropriately model them. Ultimately, the magnitude of PEs may serve as an alternative or supplementary endpoint for trials.

Similar to clinical trials, observational studies of AD provide systematic and longitudinal assessment of clinical, cognitive and pathological progression of the disease, albeit in the absence of a specific intervention. Although PEs have been relatively well studied in community-based observational studies of sporadic AD, to date, we are unaware of any systematic evaluation of PEs in the context of a clinical trial. One might expect that PEs would be attenuated in clinical trials if the study protocol includes a comprehensive screening assessment, which may provide exposure to the testing materials (Goldberg et al., 2015). Alternatively, in some cases, trial participants might be recruited from ongoing observational studies and hence are already familiar with the process of cognitive testing and may have exposure to the same test materials. Another important difference from observational studies is the role of participant expectations in clinical trials. Trial participants may exhibit enhanced PEs due to a type of placebo effect, wherein motivation and engagement may be higher in the trial compared to observational studies where expectations and motivations for participation may be different. As many trials rely on data from observational studies to select appropriate cognitive measures as endpoints and conduct power analyzes to determine the requisite sample sizes needed to detect a hypothetical treatment effect, it is critical to test the assumption that participants in observational studies will perform similarly to those engaged in clinical trial

research. If these two populations differ in terms of PEs or overall cognitive trajectories, pre-specified cognitive endpoints selected based on observational study data may not be suitable for a clinical trial and sample sizes may be underestimated, among other concerns.

To address these issues, we present analyzes from the recently completed DIAN-TU 001 (TU) clinical trial (Mills et al., 2013) and the associated DIAN Observational study (OBS, Bateman et al., 2012). The DIAN-TU is a phase 2/3, double blind, placebo controlled study of disease modifying therapies in autosomal dominant AD (ADAD), a rare form of AD due to specific genetic mutations that has similar pathological and clinical presentations, other than in age at onset, as sporadic AD (Bateman et al., 2011). These genetic mutations cause AD with virtually 100% penetrance and onset of clinical symptoms begin at a predictable and typically much younger age than sporadic AD (Ryman et al., 2014). The expected number of years to symptom onset (EYO) can be calculated based on the participant's age and the historical average age-at-symptomatic onset of gene-carriers with the same mutation or from the same family. The predictability of expected symptom onset as well as pathological similarities to the more common sporadic form of AD, makes ADAD a critical population in which to understand and build a model of cognitive, clinical, and pathological disease progression (McDade et al., 2018). To maintain participant blinding to their mutation status, ADAD mutation carriers (MCs) and non-carriers (NMCs) were enrolled in the trial, with all NMCs being assigned to placebo in a double blinded manner. The DIAN Observational study was launched in 2008 to provide natural history data on the progression of clinical, cognitive, and pathological changes in this population. Several participants who enrolled in the OBS study later enrolled in the TU study. We utilized the data from these two studies to answer the following questions: (1) Do PEs in ADAD vary as a function of mutation status or clinical status? (2) Do alternate forms that vary the stimuli across repeated administration (computerized battery vs. pen and paper) moderate the size of PEs? and (3) Do cognitive trajectories in clinical trials differ from those in observational only studies?

MATERIALS AND METHODS

A total of 384 participants were included in our analyzes. One-hundred ninety-three participants from the TU cohort and 191 from the OBS cohort. Both studies recruited a population of ADAD mutation carriers and non-carriers to determine the natural history (OBS) and to implement safe, efficient, and effective clinical trials that have the highest likelihood of success in advancing overall treatment (TU). Although the TU study was not powered to determine cognitive effects at the higher treatment doses that were ultimately used (5% power to detect a 30% slowing in the rate of cognitive decline), we have previously shown the absence of a treatment effect on cognitive outcomes in the TU (Salloway et al., 2021). Thus, given the relatively small group differences between treatment and placebo arms, for the present analyzes, all participants were combined and treatment arm [e.g., drug (solanezumab/gantenerumab) vs. placebo] was not considered. A small number of NMCs had clinical evidence for impairment (3 in the TU and 7 in OBS), these participants were removed prior to analysis due to small sample size, leaving a total of 374 participants available for analysis.

Clinical/Cognitive Evaluation

Participants in both the TU and OBS studies underwent comprehensive clinical and cognitive evaluations. Presence and severity of dementia symptoms was ascertained using the Clinical Dementia Rating® (CDR) scale (Morris, 1993). A global rating of 0 on the CDR reflects no dementia, while scores of 0.5, 1, 2, and 3 reflect very mild, mild, moderate, and severe dementia, respectively. The Mini-Mental State Exam (Folstein et al., 1975) (MMSE) was also given as a measure of general cognitive function.

The cognitive batteries were largely similar across the two studies. Neuropsychological tests that were given in common across the two cohorts have been described elsewhere (Storandt et al., 2014) and include Wechsler Memory Scale-Revised Logical Memory Immediate and Delayed Recall (Wechsler, 1987) and Digit Span, Trail making Parts A and B (Armitage, 1945), Category Fluency for Animals and Vegetables (Goodglass and Kaplan, 1983), and Digit Symbol Substitution from the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981). In the TU, participants were also administered the Cogstate computerized battery which included Identification, Detection, One-Back, One Card Learning, and the International Shopping List test. These measures have been described extensively elsewhere (Hammers et al., 2011; Lim et al., 2012). In the TU, most of these tests were administered every 6 months except for category fluency and the MMSE which were measured annually. All tests utilized the same versions at each testing occasion with the exception of the Cogstate tests which produced randomly generated stimuli at each occasion. Assessment frequency in the OBS study ranged from every 1–3 years depending on clinical status and when the participant entered the study. The OBS study has enrolled over 575 participants to date, but for the purpose of these analyzes, we selected participants that matched the enrollment criterion for the TU. We included as many participants as possible who met the following criteria: baseline global CDR score of 1 or less and estimated years to EYO range from −15 to +10 years (Salloway et al., 2021; See **Table 1** for full demographics). For the purposes of these analyzes, participants who were initially enrolled in the OBS study and then transitioned to the TU (41% of the TU CDR 0 carriers, 32% of the TU CDR > 0 carriers and 33% of the TU non-carriers started in the OBS study) were included in the TU cohort but were excluded from analyzes in the OBS cohort.

Statistical Analysis

Our analyzes proceeded in several steps. We first compared cognitive trajectories in the TU battery between NMCs, CDR 0 MCs and CDR > 0 MCs. We constructed linear mixed effects (LME) models for each cognitive test and predicted cognition from baseline EYO, time-in-study (hereafter referred to as "time"), group and the group by time interaction. A random intercept and random slope of time was also included in

TABLE 1 | Demographic characteristics of the clinical trial (TU) and observational (OBS) study cohorts.

	DIAN-TU			DIAN Obs		
	NMC	**MC CDR 0**	**MC CDR > 0**	**NMC**	**MC CDR 0**	**MC CDR > 0**
N	46	85	59	115	35	34
Age	42.0 (9.2)	40.9 (8.5)	49.2 (10.1)	41.3 (8.9)	38.7 (9.5)	46.0 (8.3)
EYO	–4.5 (6.3)	–5.8 (6.3)	2.7 (4.8)	–6.1 (6.8)	–8.3 (6.0)	1.2 (3.9)
Sex (% female)	20 (43%)	45 (53%)	28 (47%)	70 (61%)	25 (71%)	22 (65%)
Education	15.5 (3.2)	15.6 (3.2)	14.1 (2.6)	14.9 (2.8)	14.3 (2.7)	13.2 (3.2)
Number of assessments	7.3 (3.6)	9.5 (2.2)	8.1 (2.5)	2.2 (1.3)	2.7 (1.1)	3.3 (1.2)
Length of follow-up	3.1 (1.8)	4.2 (1.1)	3.6 (1.3)	2.5 (2.5)	3.6 (1.9)	2.9 (1.6)

Results are reported as mean (SD) where appropriate.

all models with an unstructured covariance matrix. Follow-up contrasts were constructed to compare slopes on each test between the NMCs and the CDR 0 MCs, and between the CDR 0 MCs and the CDR > 0 MCs. For ease of comparison across tests, all outcomes were z-scored to the baseline mean and standard deviation of the CDR 0 non-carriers so that a score of "0" represents the score of a relatively cognitively normal participant. Scores were oriented such that a positive slope indicates an improvement over time and a negative slope indicates decline.

A second set of LMEs were constructed to compare performance in the TU vs. the OBS study. Specifically, we analyzed performance on each cognitive test as a function of time, group (NMC, CDR 0 MC, and CDR > 0 MCs) and cohort (TU vs. OBS), and included all of the two and three-way interactions while also controlling for baseline EYO. All models were fit in the R statistical computing software (version 4.0.5, R Core Team, 2021) using the lme4 package (version 1.1.27.1, Bates et al., 2015). *P*-values were obtained using the lmerTest (version 3.1.3, Kuznetsova et al., 2017) package. To ensure that no influential, outlying data points were unduly biasing our results, we used the infleunce.ME package (Nieuwenhuis et al., 2012) to iteratively remove a single participant from each model and re-run the statistical analysis. We checked for a change in statistical significance in key model parameters (specifically, the group by time or group by cohort by time interactions) when a given participant was removed. Across all the analyses we conducted, none of those parameters changed significance suggesting no single person was exerting undue influence on these results. Finally, although a relatively large set of statistical comparisons were conducted in order to fully describe practice effects across a range of cognitive tests, no corrections for multiple comparisons were made.

RESULTS

Analysis 1: Clinical Trial Only

Slopes over time for each cognitive test and each group are illustrated in **Figure 1**. Intercepts and slope scores for each test can also be found in **Supplementary Table 1**. Not surprisingly, the MC CDR > 0 group evinced significant decline on all cognitive measures with some of the largest effects occurring on tests of perceptual speed and attention (Cogstate Detection, Identification and One back, Digit Symbol Substitution and Trail Making Part A). In contrast, cognitive trajectories for the MC CDR 0 group were relatively flat with a few notable exceptions. There was significant decline on Category Fluency for Animals, the ISLT and the Identification test, suggesting that measures of semantic fluency, episodic memory and attention are sensitive to preclinical cognitive decline. Interestingly, the Logical Memory immediate and delayed recall tests showed significant improvement over time in this population as did Cogstate One Card Learning, a test of visual learning ability. NMCs did not decline on any measure, which was expected in a relatively young and cognitively healthy cohort. Showing the classic pattern of practice effects, NMCs exhibited significant improvement over time compared to a zero slope on several measures including Logical Memory Immediate and Delayed Recall, Digit Symbol Substitution, Digit Span Backward and One Card Learning.

In order to determine disease effects on learning and decline, we next compared slopes between the NMCs and the MC CDR 0 group (shown in **Figure 2**) to determine if differences in rate of change distinguished the groups. Slopes (reflecting change per year in z-score units) were significantly different between these two groups on the following measures: One Card Learning (Difference = 0.10, p = 0.01, CI = 0.02:0.18), Logical Memory Immediate (Difference = 0.16, p = 0.009, CI = 0.04:0.27), Logical Memory Delayed (Difference = 0.14, p = 0.007, CI = 0.04:0.24), Digit Span Forward (Difference = 0.11, p = 0.04, CI = 0.006:0.21), Digit Span Backward (Difference = 0.13, p = 0.02, CI = 0.03:0.23) and the ISLT (Difference = 0.19, p < 0.001, CI = 0.08:0.30). These results indicate that while both MCs and NMCs exhibited PEs (see **Figure 1**) on the Logical Memory and One Card Learning tests, practice-related gains were significantly larger in the NMCs. Moreover, NMCs improved over time on the Digit Span Backward test whereas the MCs showed no significant change. Finally, the NMCs did not show improvement or decline on ISLT whereas the MCs significantly declined.

Analysis 2: Observational Versus Clinical Trial

Intercepts and slopes for the eligible participants in the OBS study are provided in **Supplementary Table 2**, and slopes for each test and group are plotted in **Figure 3**, showing time-dependent changes. First, similar to the TU, the MC CDR > 0

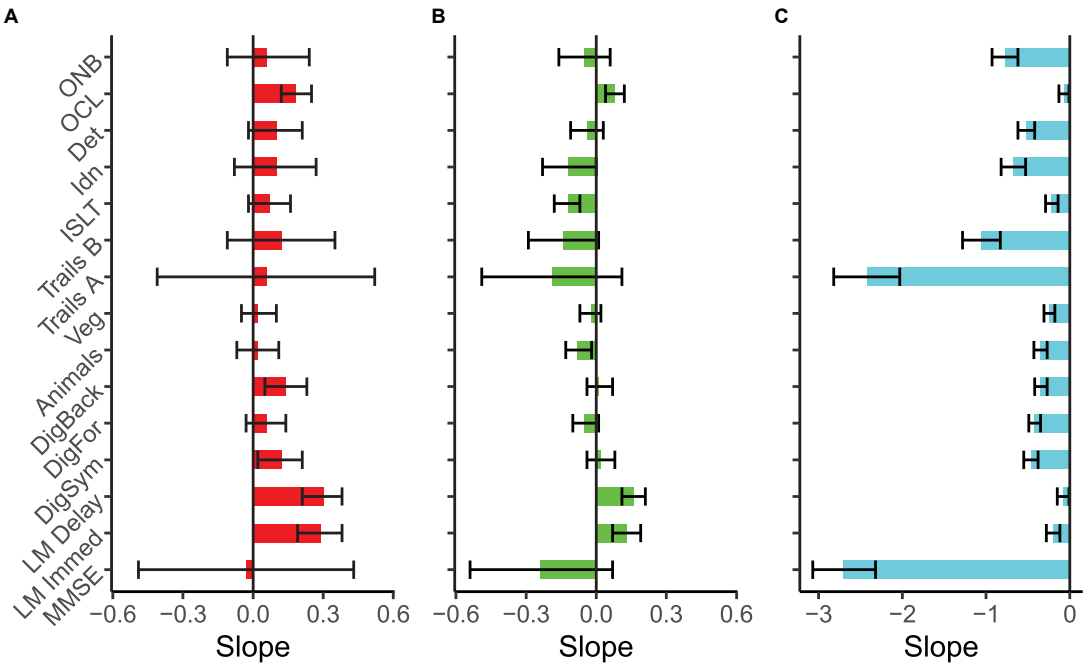


FIGURE 1 | Slope estimates and 95% confidence intervals for each test and clinical group in the Dominantly Inherited Alzheimer Network Trials Unit (DIAN-TU). Slopes can be considered significant if the CI does not encompass zero. All tests were scaled such that a negative slope indicates decline. Slopes are in z-score units change per year. **(A)** Plots non-carriers (*N* = 46), **(B)** plots Clinical Dementia Rating (CDR) 0 mutation carriers (*N* = 85), and **(C)** plots CDR > 0 mutation carriers (*N* = 59). Due to dramatic performance differences across groups, the X-axis scale is not identical across the panels. ONB, one-back; OCL, one card learning; Det, detection; Idn, identification; ISLT, international shopping list; Veg, category fluency for vegetables; animals, category fluency for animals; DigFor and DigBack, digit span forward and backward; LM, logical memory; MMSE, Mini Mental State Exam.

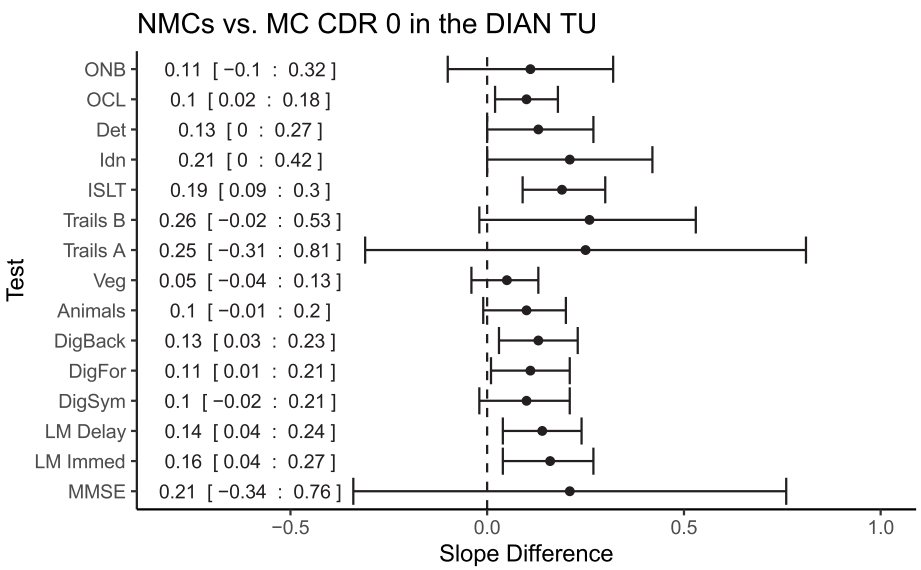


FIGURE 2 | Differences in slopes (and 95% CIs) for each test between non-carriers and CDR 0 mutation carriers in the clinical trial (TU). A positive slope difference indicates a larger or more positive slope (improvement) in the non-carriers compared to mutation carriers. Mean difference and 95% confidence intervals are pasted along the left side.

group in the observational study declined significantly on all measures. Second, the MC CDR 0 group again showed relatively flat cognitive trajectories with the notable exception of the Digit Symbol Substitution test which significantly declined by 0.12 z-score units per year. Most importantly, there was no hint of practice related improvements in the MC CDR 0s, with lack

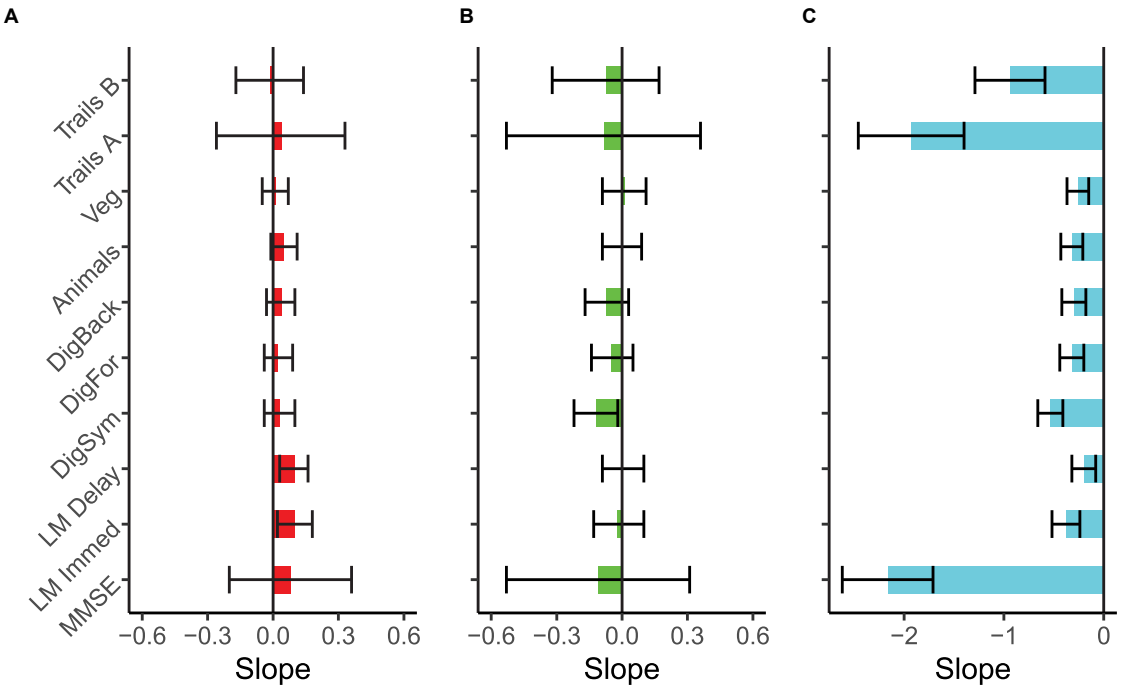


FIGURE 3 | Slope estimates and 95% confidence intervals for each test and clinical group in the DIAN-Observational (DIAN-Obs) study. Slopes are expressed as z-score units change per year and can be considered significant if the CI does not encompass zero. All tests were scaled such that a negative slope indicates decline. **(A)** Plots non-carriers (*N* = 115), **(B)** plots CDR 0 mutation carriers (*N* = 35), and **(C)** plots CDR > 0 mutation carriers (*N* = 34). Due to dramatic performance differences across groups, the X-axis scale is not identical across the panels.

of evidence of positive slope estimates, on any of the cognitive measures. Finally, the NMC group significantly improved on the Logical Memory Immediate and Delayed Recall tests but the slopes for the other measures were relatively flat and not significantly different from zero.

Direct comparisons between the symptomatic MCs in the OBS and TU cohorts (**Figure 4**), revealed no significant differences in slopes between the cohorts on any measure with the exception of Logical Memory Immediate Recall (Difference = 0.18, $p = 0.03$, $CI = 0.02{:}0.34$), in which participants in the TU showed slightly less decline than in OBS. Interestingly, a number of differences emerged when comparing the asymptomatic MCs across TU and OBS (**Figure 5**). Specifically, on the Digit Symbol Substitution test (Difference = 0.14, $p = 0.02$, $CI = 0.02{:}0.25$), Logical Memory Immediate (Difference = 0.15, $p = 0.03$, $CI = 0.02{:}0.28$) and Delayed recall (Difference = 0.15, $p = 0.007$, $CI = 0.04{:}0.26$) slopes were markedly less negative in the TU as compared to the OBS study. Finally, in the comparison of NMCs (**Figure 6**), the OBS participants improved less on Logical Memory Immediate (Difference = 0.19, $p = 0.003$, $CI = 0.07{:}0.31$) and Delayed recall (Difference = 0.20, $p < 0.001$, $CI = 0.10{:}0.30$) compared to the TU participants.

DISCUSSION

In this study, we compared performance on a comprehensive cognitive battery in two cohorts to answer several important questions regarding practice related improvements in observational studies and clinical trials in AD populations.

Question 1: Does Mutation Status or Clinical Status Moderate Practice Effects in the Dominantly Inherited Alzheimer Network Trials Unit?

Clinical status was an important predictor of PEs in the DIAN-TU. Specifically, individuals who were CDR > 0 at entry significantly declined on all cognitive measures and therefore did not show practice-related gains. This is not to say that PEs were not present in this group, only that any gains associated with practice were overshadowed by the decline attributable to AD pathology. More importantly, mutation status in the CDR 0 groups also predicted magnitude of change in the TU. MC CDR 0s declined significantly over time on measures of attention, episodic memory, and semantic fluency whereas NMCs showed no change in these domains. Interestingly, differences in performance between MC CDR 0s and NMCs also emerged on the ISLT (list recall, MCs declined more than NMCs), Logical Memory (narrative recall, MCs improved less than NMCs), Digit Span (working memory, MCs improved less than NMCs), and One Card Learning (visual learning, MCs improved less than NMCs). Together these findings suggest that differences in the magnitude of practice related improvements in domains of memory and learning might serve as a sensitive and supplemental indicator of preclinical AD.

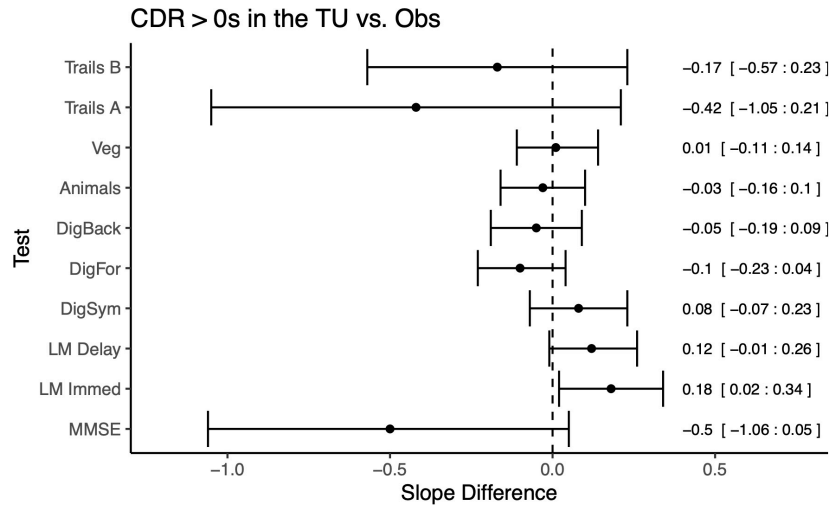


FIGURE 4 | Differences in slopes (and 95% CIs) for each test between CDR > 0 mutation carriers in TU vs. the Obs study. A positive slope difference indicates a larger or more positive slope (improvement) in the TU compared to the Obs study. Some tests, such as Logical Memory, had more improvement or practice effects in the TU vs. Obs. Mean differences and 95% CIs presented along the right side of the graph.

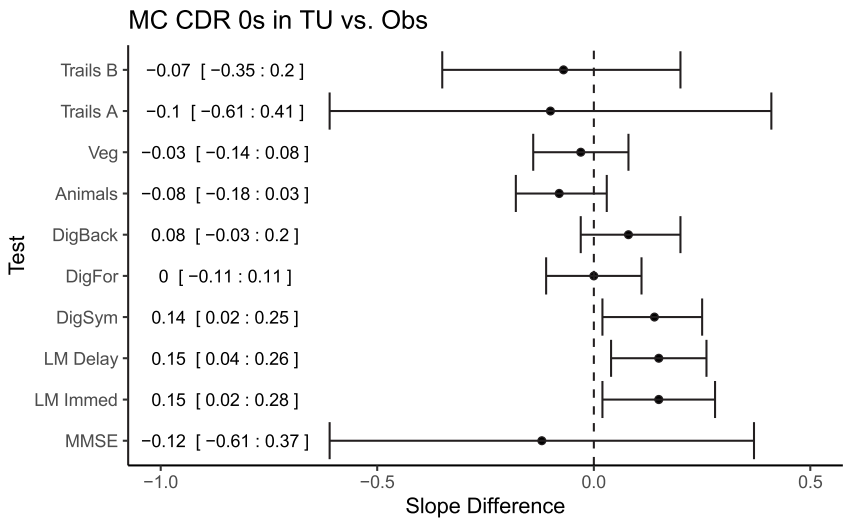


FIGURE 5 | Differences in slopes (and 95% CIs) for each test between CDR 0 mutation carriers in TU vs. the Obs study. A positive slope difference indicates a larger or more positive slope (improvement) in the TU compared to the Obs study. Some tests, such as Logical Memory and Digit Symbol, had more improvement or practice effects in the TU vs. Obs. Mean differences and 95% CIs presented along the left side of the graph.

Question 2: Do Alternative Forms Influence Practice Effects?

We expected *a priori* that computerized measures from the Cogstate battery might show less practice effects due to the nature of randomized stimuli which generates essentially unlimited alternate forms. For example, many of these tasks use playing cards as stimuli presented in a newly randomized order at each administration. Such a design reduces the possibility of memorizing specific items which can be a contributor to PEs. This contrasts with Logical Memory in the DIAN studies, for example, which presents the same narrative each time the test is taken.

Our hypothesis was largely supported. Most of the computerized tests were resistant to practice effects in the NMCs or sensitive to decline in the MCs (e.g., ISLT and Identification tests). Practice related gains were apparent on the One Card Learning test and due to the nature of the randomized stimuli, it is assumed that participants are developing or learning some strategy besides rote memorization to improve over time. One possibility is that this test might be particularly amenable to visual strategies such as the method of loci (Gross et al., 2014). As the cards are shown one at a time, participants may over time learn to organize the items in a meaningful fashion (e.g., into poker hands or by suit) which might aid recall.

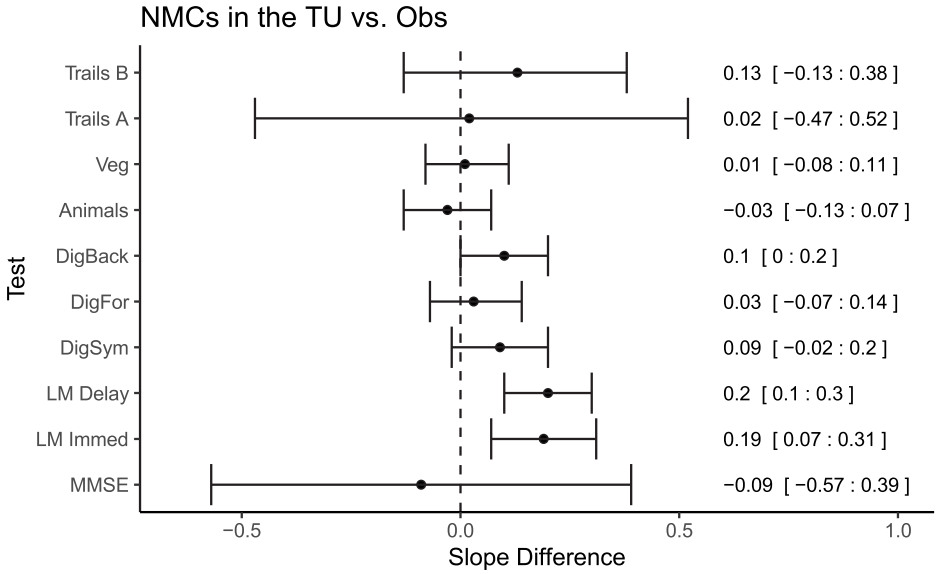


FIGURE 6 | Differences in slopes (and 95% CIs) for each test between non-mutation carriers in TU vs. the Obs study. A positive slope difference indicates a larger or more positive slope (improvement) in the TU compared to the Obs study. Some tests, such as Logical Memory had more improvement or practice effects in the TU vs. Obs. Mean differences and 95% CIs presented along the right side of the graph.

Question 3: Are Practice Effects Similar Across Clinical Trials and Observational Studies?

One of the most important questions addressed in this study was whether cognitive trajectories were similar across a clinical trial cohort and an observational study. For participants who were CDR > 0 at baseline, the answer was clearly "yes". Regardless of the cohort, MC CDR > 0s declined significantly over time and the magnitude of change did not differ significantly between the TU and OBS with the sole exception of Logical Memory Immediate Recall. This may reflect disease progression such that symptomatic MCs have declined to the extent that any practice related gains were outweighed by the task demands. An interesting but complex question for future studies is to determine the point at which PEs are effectively overwhelmed by disease related declines.

For the MC CDR 0s, however, a few critical differences did emerge. Specifically, OBS participants declined at a faster rate than the TU participants on the Digit Symbol Substitution test and improved less on the Logical Memory Immediate and Delayed Recall tests. One obvious possible explanation for these differences is the assessment frequency across the two studies (every 6 months in the TU, ∼ every 2 years in OBS). This explanation is likely for the Logical Memory tests, where participants will hear the same story at each testing occasion which reinforces encoding and aids in recall. It is less clear why Digit Symbol Substitution would show such enhanced practice effects in the TU when other measures of speed and executive function did not (e.g., the Trail Making tests). Studies of retest have shown performance gains on Digit Symbol Substitution, but this test typically demonstrates less gains than episodic memory measures (Calamia et al., 2012). Thus, frequency of assessment needs to be carefully considered during trial design.

Another important possibility is an enhanced placebo effect in the DIAN-TU. Specifically, TU participants were randomized to treatment vs. placebo at a ratio of 3:1. Thus, there may have been a greater expectation of being on active drug which may have then impacted cognitive performance. Regardless of the underlying mechanisms, these differences in practice related gains are particularly noteworthy as the Logical Memory and Digit Symbol tests feature heavily in multiple cognitive composite endpoints (Sperling et al., 2014; Bateman et al., 2017). Investigators should keep in mind potential differences between observational and trial cohorts when planning their studies and conducting power analyzes.

Using alternate instrument forms has been shown in some studies to be a viable strategy to reduce PEs. For example, a meta-analysis of test/retest effects found substantial reductions in performance gains when alternate forms were used for verbal list learning measures (Calamia et al., 2012). This finding is similar to the results shown here, in which the computerized tests were largely resistant to practice-related gains. The one exception we found was One Card Learning, a visual learning test that uses randomly generated sequences of cards such that there are essentially hundreds of alternate forms. This task produced the largest PEs in asymptomatic MCs enrolled in the DIAN-TU clinical trial. We could not, however, determine if this was due solely to clinical trial participation, as this measure was not collected with sufficient samples in the OBS study for comparison. In a recent study, the developers of the One Card Learning test made a shorter and less difficult version of the test (as evidenced by less floor effects in symptomatic AD participants) that demonstrated no PEs in young cognitively

normal participants across very short retest intervals (White et al., 2021). The authors argue that the increased difficulty and length of the longer version of the task may lead to participants forming strategies that in turn lead to more PEs.

Although our results indicate that rates of change on key cognitive outcomes may be underestimated in clinical trials due to the presence of these practice effects, it is important to highlight situations in which these practice effects might limit the ability to detect treatment effects. Specifically, in clinical trials that include a placebo arm in which participants undergo identical clinical and cognitive assessments as participants on active treatment, the negative impact of practice effects may be minimal, to the extent that practice effects manifest similarly in placebo vs. treated patients. However, this also assumes that the influence of improved cognition due to treatment is additive, rather than interactive, with improved cognition due to practice effects, which may not be the case. Moreover, the primary cognitive outcome is often a composite score formed of multiple tests. If some tests exhibit practice effects while others do not, as is the case in the present study, decline on a global composite score may be very small, limiting the power to detect any differences among groups.

It is unclear if attempts to avoid or reduce practice effects are futile. Completely avoiding practice effects does seem an impossible task. One of the most fundamental aspects of human behavior is adaptation, or learning. As we and others have previously shown, in the context of a cognitive assessment this learning is not just limited to familiarity with test materials but also to process factors like test strategies, effort, demand characteristics, and expectancy effects, among others (Beglinger et al., 2005; Hassenstab et al., 2015; Machulda et al., 2017). Instead of avoiding PEs, trials that enroll cognitively normal or mildly affected participants might consider designs and statistical models that anticipate and account for the influence of PEs. Such protocols might include extended baseline designs that allow cluster assessments prior to dosing in so-called "run-in" designs (Frost et al., 2008). Less emphasis might be placed on spreading assessments out at regular time intervals (e.g., one assessment every 6 months) in favor of clustering assessments at key read-out times and averaging across the clusters, which might not only minimize the effects of practice but also reduce individual variability in scores (Valdes et al., 2016). An alternative strategy is to incorporate PEs as outcomes themselves. Several recent studies have deliberately measured learning effects in cognitively normal older adults at risk for AD (Hassenstab et al., 2015; Baker et al., 2020; Lim et al., 2020; Samaroo et al., 2020). Effect sizes differentiating participants with biomarker-confirmed preclinical AD from those with normal biomarker levels are extraordinarily large for these paradigms, suggesting that PEs may be a highly sensitive indicator of disease progression.

There are many strengths to this study including use of a comprehensive cognitive battery on very well-characterized clinical cohorts, designed comparability between an observational study and clinical trial, enrolling the same population for both studies, and frequent assessments over many years. However, some limitations need to be noted. First, because this is a study of ADAD, a very rare form of AD, the sample sizes included here could be considered small. Moreover, it is unclear whether differences in practice related gains will translate to the more common sporadic form of the disease. Second, some participants in these studies may become aware of their mutation status and this might alter their cognitive outcomes (Aschenbrenner et al., 2020). It is unknown whether the number of participants who did and did not learn their status were similar across the two studies. Third, we did not have data from the Cogstate testing battery in the DIAN-OBS study which precluded a comparison of PEs between the trial and observational study on these measures. Finally, we conducted many statistical tests due to the large cognitive battery that was administered and although many effects could have been predicted *a priori* this could be seen as an additional limitation.

Nevertheless, these results highlight three important points. (1) Practice effects were highly evident in the DIAN-TU-001 clinical trial in asymptomatic mutation carriers and non-carriers. (2) Alternate forms may have attenuated practice effects, but not for all measures. (3) The magnitudes of practice effects were larger in the DIAN-TU-001 clinical trial than seen in a well-matched sample from the DIAN Observational study, suggesting that more frequent assessments and placebo effects in clinical trials may drive increases in practice effects. Clinical trials that utilize a cognitive endpoint should carefully consider the potential for practice effects and select statistical modeling strategies that can incorporate them directly.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because risk of identifying individual participants and/or risk to ongoing trial activities. Requests to access the datasets should be directed to https://dian.wustl.edu/our-research/for-investigators/diantu-investigator-resources.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Local Ethics Committees at TU sites. The patients/participants provided their written informed consent to participate in this study.

DOMINANTLY INHERITED ALZHEIMER NETWORK TRIALS UNIT (DIAN-TU)

Data used in the preparation of this article were obtained from the Dominantly Inherited Alzheimer Network Trials Unit (DIAN-TU). As such, the study team members within the DIAN-TU contributed to the design and implementation of DIAN-TU and/or provided data but may not have participated in the analysis or writing of this report. A complete listing of the DIAN-TU Study Team Members can be found at dian.wustl.edu, DIAN-TU Study Team.

AUTHOR CONTRIBUTIONS

AA statistical analysis, development of study hypotheses, and initial draft of the manuscript. JH development of study hypotheses, acquisition of study data, and revised manuscript for intellectual content. GW, YL, CX, EC, and KH revised manuscript for intellectual content. EM, DC, SS, MF, RY, CJM, CLM, G-YH, GS, GD, SW, LH, JG, JR, WB, NF, PS, KS, HS, SG, and RB acquisition of study data and revised manuscript for intellectual content. All authors contributed to the article and approved the submitted version.

FUNDING

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under award numbers U01AG042791 and U01AG042791-S1 (FNIH and Accelerating Medicines Partnership), R1AG046179 and R01AG053267-S1. The research for the DIAN-TU-001 trial, solanezumab and gantenerumab drug arms was also supported by the Alzheimer's Association, Eli Lilly and Company, F. Hoffman-LaRoche Ltd., Avid Radiopharmaceuticals (a wholly owned subsidiary of Eli Lilly and Company), GHR Foundation, an anonymous organization, Cogstate, and Signant. The DIAN-TU has received funding from the DIAN-TU Pharma Consortium. We acknowledged the altruism of the participants and their families and contributions of the DIAN, DIAN Expanded Registry, and DIAN-TU research and support staff at each of the participating sites (see DIAN-TU Study Team) for their contributions to this study.

ACKNOWLEDGMENTS

This manuscript has been reviewed by DIAN-TU Study investigators for scientific content and consistency of data interpretation with previous DIAN-TU Study publications. David Holtzman, former Department Head of Neurology where the research was conducted, an inventor on patents for one of the treatments (solanezumab), which has been tested in the DIAN-TU clinical trials. If solanezumab is approved as a treatment for Alzheimer's disease or Dominantly Inherited Alzheimer's Disease, Washington University and Holtzman will receive part of the net sales of solanezumab from Eli Lilly, which has licensed the patents related to solanezumab from Washington University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.883131/full#supplementary-material

REFERENCES

Armitage, S. G. (1945). An analysis of certain psychological tests used for the evaluation of brain injury. *Psychol. Monogr.* 60, 1–48.

Aschenbrenner, A. J., James, B. D., McDade, E., Wang, G., Lim, Y. Y., Benzinger, T. L. S., et al. (2020). Awareness of genetic risk in the Dominantly Inherited Alzheimer Network (DIAN). *Alzheimers Dement.* 16, 219–228. doi: 10.1002/alz.12010

Baker, J. E., Bruns, L., Hassenstab, J., Masters, C. L., Maruff, P., and Lim, Y. Y. (2020). Use of an experimental language acquisition paradigm for standardized neuropsychological assessment of learning: a pilot study in young and older adults. *J. Clin. Exp. Neuropsychol.* 42, 55–65. doi: 10.1080/13803395.2019.1665626

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., and Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neurosci.* 11:118. doi: 10.1186/1471-2202-11-118

Bateman, R. J., Aisen, P. S., De Strooper, B., Fox, N. C., Lemere, C. A., Ringman, J. M., et al. (2011). Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimers Res. Ther.* 3:1. doi: 10.1186/alzrt59

Bateman, R. J., Benzinger, T. L., Berry, S., Clifford, D. B., Duggan, C., Fagan, A. M., et al. (2017). The DIAN-TU Next Generation Alzheimer's prevention trial: adaptive design and disease progression model. *Alzheimers Dement.* 13, 8–19. doi: 10.1016/j.jalz.2016.07.005

Bateman, R. J., Xiong, C., Benzinger, T. L., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and biomarker changes in Dominantly Inherited Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beglinger, L., Gaydos, B., Tangphaodaniels, O., Duff, K., Kareken, D., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch. Clin. Neuropsychol.* 20, 517–529. doi: 10.1016/j.acn.2004.12.003

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Collie, A., Maruff, P., Darby, D. G., and McSTEPHEN, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *J. Int. Neuropsychol. Soc.* 9, 419–428. doi: 10.1017/S1355617703930074

Cummings, J., Lee, G., Ritter, A., Sabbagh, M., and Zhong, K. (2020). Alzheimer's disease drug development pipeline: 2020. *Alzheimers Dement. Transl. Res. Clin. Interv.* 6:e12050. doi: 10.1002/trc2.12050

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., et al. (2011). Practice Effects Predict Cognitive Outcome in Amnestic Mild Cognitive Impairment. *Am. J. Geriatr. Psychiatry* 19, 932–939. doi: 10.1097/JGP.0b013e318209dd3a

Falleti, M. G., Maruff, P., Collie, A., and Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the cogstate battery at 10-minute, one week and one month test-retest intervals. *J. Clin. Exp. Neuropsychol.* 28, 1095–1112. doi: 10.1080/13803390500205718

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198.

Frost, C., Kenward, M. G., and Fox, N. C. (2008). Optimizing the design of clinical trials where the outcome is a rate. Can estimating a baseline rate in a run-in period increase efficiency? *Stat. Med.* 27, 3717–3731. doi: 10.1002/sim.3280

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement.* 1, 103–111. doi: 10.1016/j.dadm.2014.11.003

Goodglass, H., and Kaplan, E. (1983). *Boston Diagnostic AphasiaExamination Booklet, III, ORAL EXPRESSION, J. Animal Naming (Fluency in Controlled Association*. Philadelphia: Lea & Febiger.

Gross, A. L., Brandt, J., Bandeen-Roche, K., Carlson, M. C., Stuart, E. A., Marsiske, M., et al. (2014). Do older adults use the method of loci? Results from the

ACTIVE study. *Exp. Aging Res.* 40, 140–163. doi: 10.1080/0361073X.2014.882204

Gross, A. L., Inouye, S., Rebok, G., Brandt, J., Crane, P. K., Parisi, J., et al. (2012). Parallel but not equivalent: challenges and solutions for repeated assessment of cognition over time. *J. Clin. Exp. Neuropsychol.* 34, 758–772. doi: 10.1080/13803395.2012.681628

Hammers, D., Spurgeon, E., Ryan, K., Persad, C., Heidebrink, J., Barbas, N., et al. (2011). Reliability of repeated cognitive assessment of dementia using a brief computerized battery. *Am. J. Alzheimers Dis. Dementiasr.* 26, 326–333. doi: 10.1177/1533317511411907

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Jacobs, D. M., Ard, M. C., Salmon, D. P., Galasko, D. R., Bondi, M. W., and Edland, S. D. (2017). Potential implications of practice effects in Alzheimer's disease prevention trials. *Alzheimers Dement. Transl. Res. Clin. Interv.* 3, 531–535. doi: 10.1016/j.trci.2017.08.010

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). **lmerTest** Package: tests in Linear Mixed Effects Models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lim, Y. Y., Baker, J. E., Bruns, L., Mills, A., Fowler, C., Fripp, J., et al. (2020). Association of deficits in short-term learning and Aβ and hippocampal volume in cognitively normal adults. *Neurology* 95, e2577–e2585. doi: 10.1212/WNL.0000000000010728

Lim, Y. Y., Harrington, K., Ames, D., Ellis, K. A., Lachovitzki, R., Snyder, P. J., et al. (2012). Short term stability of verbal memory impairment in mild cognitive impairment and Alzheimer's disease measured using the International Shopping List Test. *J. Clin. Exp. Neuropsychol.* 34, 853–863. doi: 10.1080/13803395.2012.689815

Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., et al. (2017). Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. *Clin. Neuropsychol.* 31, 99–117. doi: 10.1080/13854046.2016.1241303

McDade, E., Wang, G., Gordon, B. A., Hassenstab, J., Benzinger, T. L. S., Buckles, V., et al. (2018). Longitudinal cognitive and biomarker changes in dominantly inherited Alzheimer disease. *Neurology* 91, e1295–e1306. doi: 10.1212/WNL.0000000000006277

Mills, S. M., Mallmann, J., Santacruz, A. M., Fuqua, A., Carril, M., Aisen, P. S., et al. (2013). Preclinical trials in autosomal dominant AD: implementation of the DIAN-TU trial. *Rev. Neurol.* 169, 737–743. doi: 10.1016/j.neurol.2013.07.017

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* 43, 2412–2414. doi: 10.1212/wnl.43.11.2412-a

Nieuwenhuis, R., Te Grotenhuis, M., and Pelzer, B. (2012). influence.ME: tools for detecting influential data in mixed effects models. *R J.* 4, 38–47.

Oltra-Cucarella, J., Sánchez-SanSegundo, M., and Ferrer-Cascales, R. (2018). Cognition or genetics? Predicting Alzheimer's disease with practice effects, APOE genotype, and brain metabolism. *Neurobiol. Aging* 71, 234–240. doi: 10.1016/j.neurobiolaging.2018.08.004

R Core Team (2021). *R: A Language and Environment for Statisical Computing*. Vienna: R Foundation for Statistical Computing.

Ryman, D. C., Acosta-Baena, N., Aisen, P. S., Bird, T., Danek, A., Fox, N. C., et al. (2014). Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. *Neurology* 83, 253–260. doi: 10.1212/WNL.0000000000000596

Salloway, S., Farlow, M., McDade, E., Clifford, D. B., Wang, G., Llibre-Guerra, J. J., et al. (2021). A trial of gantenerumab or solanezumab in dominantly inherited Alzheimer's disease. *Nat. Med.* 27, 1187–1196. doi: 10.1038/s41591-021-01369-8

Salthouse, T. A., Schroeder, D. H., and Ferrer, E. (2004). Estimating Retest Effects in Longitudinal Assessments of Cognitive Functioning in Adults Between 18 and 60 Years of Age. *Dev. Psychol.* 40, 813–822. doi: 10.1037/0012-1649.40.5.813

Samaroo, A., Amariglio, R. E., Burnham, S., Sparks, P., Properzi, M., Schultz, A. P., et al. (2020). Diminished Learning Over Repeated Exposures (LORE) in preclinical Alzheimer's disease. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* 12:e12132. doi: 10.1002/dad2.12132

Sperling, R. A., Rentz, D. M., Johnson, K. A., Karlawish, J., Donohue, M., Salmon, D. P., et al. (2014). The A4 Study: Stopping AD Before Symptoms Begin? *Sci. Transl. Med.* 6, fs13–fs228. doi: 10.1126/scitranslmed.3007941

Storandt, M., Balota, D. A., Aschenbrenner, A. J., and Morris, J. C. (2014). Clinical and psychological characteristics of the initial cohort of the dominantly inherited Alzheimer Network (DIAN). *Neuropsychology* 28:19. doi: 10.1037/neu0000030

Valdes, E. G., Sadeq, N. A., Harrison Bush, A. L., Morgan, D., and Andel, R. (2016). Regular cognitive self-monitoring in community-dwelling older adults using an internet-based tool. *J. Clin. Exp. Neuropsychol.* 38, 1026–1037. doi: 10.1080/13803395.2016.1186155

Wechsler, D. (1981). *Manual: Wechsler Adult Intelligence Scale- Revised*. New York, NY: Psychological Corporation.

Wechsler, D. (1987). *Manual: Wechsler Memory Scale- Revised*. San Antonio, TX: Psychological Corporation.

White, J. P., Schembri, A., Edgar, C. J., Lim, Y. Y., Masters, C. L., and Maruff, P. (2021). A paradox in digital memory assessment: increased sensitivity with reduced difficulty. *Front. Digit. Health* 3:780303. doi: 10.3389/fdgth.2021.780303

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: EC, KH, and RY are full-time employees and/or stockholders of Eli Lilly. RB has received funding from Avid Radiopharmaceuticals, Jansse, Hoffman La-Roche/Genentech, Eli LIlly & Co., Eisai, Biogen, AbbVie, and Bristol Meyer Squibbs, has royalties/licenses from C2N Diagnostics, consulting fees from Eisai, Amgen, and Hoffman La-Roche, honoraria from the Korean Dementia Association and the American Neurological Association, and is on the data safety monitoring board or advisory board for Roche/Genentech and Biogen. Unrelated to this article, RB serves as the principal investigator of the DIAN-TU, which is supported by the Alzheimer's Association, GHR Foundation, an anonymous organization, and the DIAN-TU Pharma Consortium (Active: Eli Lilly and Company/Avid Radiopharmaceuticals, F. Hoffman-La Roche/Genentech, Biogen, Eisai, and Janssen. Previous: Abbvie, Amgen, AstraZeneca, Forum, Mithridion, Novartis, Pfizer, Sanofi, and United Neuroscience). In addition, in-kind support has been received from CogState and Signant Health. G-YH has received research support as a clinical trials site investigator from Anavax, Biogen, Eli Lilly, and Roche, has received research grants from the CIHR, Alzheimer Society of Canada, and NIA/NIH, is supported by the Ralph Fisher Professorship in dementia research from the Alzheimer Society of British Columbia, and has participated in expert advisory committees sponsored by Biogen and Roche. GD's research is supported by NIH (K23AG064029), the Alzheimer's Association, and Chan Zuckerberg Initiative, and he serves as a consultant for Parabon Nanolabs Inc., as a Topic Editor (Dementia) for DynaMed (EBSCO), and as the Clinical Director of the Anti-NMDA Receptor Encephalitis Foundation (Inc, Canada; uncompensated) and owns stock in ANI pharmaceuticals. DH, former Department Head of Neurology where the research was conducted, is an inventor on patents for one of the treatments (solanezumab), which has been tested in the DIAN-TU clinical trials. If solanezumab is approved as a treatment for Alzheimer's disease or Dominantly Inherited Alzheimer's Disease, Washington University, and will receive part of the net sales of solanezumab from Eli Lilly, which has licensed the patents related to solanezumab from Washington University.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 06 July 2022
doi: 10.3389/fnagi.2022.909614

Practice Effect of Repeated Cognitive Tests Among Older Adults: Associations With Brain Amyloid Pathology and Other Influencing Factors

Bang Zheng[1,2], Chinedu Udeh-Momoh[1], Tamlyn Watermeyer[3], Celeste A. de Jager Loots[1], Jamie K. Ford[1], Catherine E. Robb[1], Parthenia Giannakopoulou[1], Sara Ahmadi-Abhari[1], Susan Baker[4], Gerald P. Novak[4], Geraint Price[1] and Lefkos T. Middleton[1,5]*

[1] Ageing Epidemiology (AGE) Research Unit, School of Public Health, Imperial College London, London, United Kingdom, [2] Department of Non-communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, [3] Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom, [4] Janssen Research and Development LLC, Titusville, NJ, United States, [5] Public Health Directorate, Imperial College NHS Healthcare Trust, London, United Kingdom

OPEN ACCESS

Edited by:
Daniel Nation,
University of California, Irvine,
United States

Reviewed by:
Matthew Grilli,
University of Arizona, United States
Alexandra Weigand,
University of California, San Diego,
United States

***Correspondence:**
Geraint Price
g.price@imperial.ac.uk

Specialty section:
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

Received: 31 March 2022
Accepted: 16 June 2022
Published: 06 July 2022

Citation:
Zheng B, Udeh-Momoh C, Watermeyer T, de Jager Loots CA, Ford JK, Robb CE, Giannakopoulou P, Ahmadi-Abhari S, Baker S, Novak GP, Price G and Middleton LT (2022) Practice Effect of Repeated Cognitive Tests Among Older Adults: Associations With Brain Amyloid Pathology and Other Influencing Factors. Front. Aging Neurosci. 14:909614. doi: 10.3389/fnagi.2022.909614

Background: Practice effects (PE), after repeated cognitive measurements, may mask cognitive decline and represent a challenge in clinical and research settings. However, an attenuated practice effect may indicate the presence of brain pathologies. This study aimed to evaluate practice effects on the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) scale, and their associations with brain amyloid status and other factors in a cohort of cognitively unimpaired older adults enrolled in the CHARIOT-PRO SubStudy.

Materials and Methods: 502 cognitively unimpaired participants aged 60-85 years were assessed with RBANS in both screening and baseline clinic visits using alternate versions (median time gap of 3.5 months). We tested PE based on differences between test and retest scores in total scale and domain-specific indices. Multiple linear regressions were used to examine factors influencing PE, after adjusting for age, sex, education level, *APOE*-ε4 carriage and initial RBANS score. The latter and PE were also evaluated as predictors for amyloid positivity status based on defined thresholds, using logistic regression.

Results: Participants' total scale, immediate memory and delayed memory indices were significantly higher in the second test than in the initial test (Cohen's d_z = 0.48, 0.70 and 0.35, $P < 0.001$). On the immediate memory index, the PE was significantly lower in the amyloid positive group than the amyloid negative group ($P = 0.022$). Older participants

(≥70 years), women, non-*APOE*-ε4 carriers, and those with worse initial RBANS test performance had larger PE. No associations were found between brain MRI parameters and PE. In addition, attenuated practice effects in immediate or delayed memory index were independent predictors for amyloid positivity ($P < 0.05$).

Conclusion: Significant practice effects on RBANS total scale and memory indices were identified in cognitively unimpaired older adults. The association with amyloid status suggests that practice effects are not simply a source of measurement error but may be informative with regard to underlying neuropathology.

Keywords: practice effect, cognitive test, older adults, amyloid pathology, memory

INTRODUCTION

Valid instruments and implementations of cognitive tests are essential for the evaluation of cognitive status, decline and subsequent dementia diagnosis, and the screening of at-risk participants for clinical trials and population intervention programs for dementia prevention. However, practice effects (PE) after repeated cognitive measurements, which refer to improvements in test performance due to repeated exposure to test materials or procedures (Hausknecht et al., 2007; Goldberg et al., 2015), often mask a potential cognitive decline and remain a major issue in clinical and research settings (Houx et al., 2002; Sanderson-Cimino et al., 2022). Failing to account for practice effects in cognitive tests could delay diagnosis and clinical care for patients with cognitive deficits. PE resulting from task familiarity occurring with test repetition is distinct from learning effects which refer to the recall of correct answers from previous tests. The latter is often addressed in neuropsychological practice through administration of alternate versions of the same task (e.g., different word lists in verbal memory tests).

Exploring factors that influence practice effects can be informative of potential heterogeneity of measurement bias and in developing mitigation strategies to minimise such bias (Calamia et al., 2012). On the other hand, the magnitude of practice effect *per se* may also have indicative value for cognitive impairment or existing brain pathologies (Duff et al., 2007; Jutten et al., 2021). From this perspective, PE may represent not merely a source of measurement error but potentially valuable information from a clinical and scientific perspective (Duff et al., 2007).

Given the long preclinical stage of late-onset dementia (Elias et al., 2000) with progressively accumulating neuropathology, it is early detection in at-risk individuals that may prove essential in reducing the burden of cognitive and functional decline and dementia in the elderly population. Therefore, a deeper understanding and characterisation of PE in validated cognitive assessment tools among asymptomatic population is warranted.

This study aimed to evaluate PE in the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) (Randolph et al., 1998), and its associations with brain amyloid status and other factors in a cohort of cognitively unimpaired older adults in the United Kingdom Cognitive Health in Ageing Register: Investigational, Observational, and Trial Studies in Dementia Research: Prospective Readiness cOhort Study (CHARIOT-PRO) SubStudy (Udeh-Momoh et al., 2021).

MATERIALS AND METHODS

Study Population

CHARIOT-PRO SubStudy is an on-going prospective cohort study of cognitively unimpaired older adults in the United Kingdom, which aims to examine longitudinal cognitive changes in those with and without brain amyloid-beta (Aβ42) pathology, and factors and markers of subsequent decline (Udeh-Momoh et al., 2021). Following screening of 2425 individuals, including amyloid status determination and multiple cognitive tests, an equal number of participants above and below a binary threshold of Aβ42 positivity were enrolled at baseline and in subsequent longitudinal study. During screening, participants whose performance on any RBANS index was poorer than 1.5 standard deviation (SD) below the population mean (population norms from Randolph, 1998) were referred to an adjudication panel of neurologists, psychiatrists and neuropsychologists to detect any undiagnosed cognitive impairment which was an exclusion criterion. The detailed inclusion/exclusion criteria and study procedures have been described in previous papers of our group (Nalder et al., 2021; Udeh-Momoh et al., 2021). The study received approval from the National Research Ethics Service (NRES) Committee London Central [reference 15/LO/0711 (IRAS 140764)], as well as independent ethics review by committees from the local sites. All participants provided informed consent before participating in the study.

A total of 502 participants aged 60–85 years completed RBANS assessments in both screening and baseline clinic visits and were included in this study (Udeh-Momoh et al., 2021). The median time gap between the screening visit and the baseline visit was 3.5 months, which allowed us to examine the practice effects in RBANS scale within a relatively short time period with less concern that the test-retest score differences are (partially) due to the cognitive decline during this time interval.

Measurements

Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) (Randolph et al., 1998) is a validated and widely used neuropsychological assessment. It is a 20-min composite battery which consists of twelve subtests that measure five cognitive domain indices (immediate memory, delayed memory, visuospatial construction, language, attention). The sum of the five index scores is converted to a total scale score based on a

distribution with a mean of 100 and SD of 15. This assessment was administered by trained assistant psychologists during the in-person clinic visits. Version C and Version A of the RBANS were administered at the screening and baseline assessments, respectively, to avoid learning effects (i.e., recalling answers from the same test received before).

Amyloid burden was determined during the screening visit either by amyloid positron emission tomography (PET) scans (in ∼90% of participants) or cerebrospinal fluid (CSF) Aβ42 measurements via lumbar punctures (in the remaining 10%). Aβ positive was defined as above-threshold brain Aβ deposition on PET (based on tracer-specific thresholds of the composite cortical standardised uptake value ratio, SUVR) or below-threshold CSF Aβ42 concentration (≤600 ng/L). Three F18-radiolabeled amyloid tracers were used: florbetapir (Amyvid), flutemetamol (Vizamyl) and florbetaben (Neuraceq). The composite cortical SUVR threshold was 1.14 for Amyvid and 1.23 for Vizamyl (both with whole cerebellum as reference region), and 1.20 for Neuraceq (with cerebellar grey matter as reference region) (Udeh-Momoh et al., 2021).

Screening also included a brain magnetic resonance imaging (MRI). Bilateral volumetric MRI parameters were obtained, including whole brain volume (mL^3), ventricular volume (mL^3), hippocampal volume (mm^3) and AD signature cortical thickness (mm) (Schwarz et al., 2016). Intracranial volume (ICV) was used as the proxy variable for premorbid brain volume to be adjusted for in the analyses of MRI parameters. All study procedures and cut-off points have previously been reported (Udeh-Momoh et al., 2021).

We also collected other information including age, sex, ethnicity, education level, *APOE* genotype and National Adult Reading Test (NART) score [as a proxy for premorbid intelligence quotient (IQ)] (Nelson and Willison, 1991).

Statistical Analyses

Demographic and clinical characteristics of study participants were compared according to amyloid pathology status (amyloid positive vs. negative) using independent samples *t*-test, chi-squared test, rank-sum test or general linear regression, where appropriate. We assessed the internal consistency reliability (Cronbach's α coefficient) and test-retest reliability (Pearson correlation coefficient *r*) of the RBANS scale in this cohort. PE was estimated based on differences between test and retest scores (i.e., measurements at the screening and baseline visits) in RBANS total scale and domain-specific indices. Paired *t*-test was used to test the statistical significance of PE; Cohen's d_z for the within-subjects design (Cohen, 1988) was calculated as the standardised effect size for PE (i.e., scaled difference scores).

Multiple linear regression model was used to examine whether the magnitude of PE varies by amyloid status, with the test-retest difference score in RBANS total scale or domain-specific index as the dependent variable, amyloid status as the independent variable of interest, while adjusting for age, sex, education level, *APOE*-ε4 carriage and initial RBANS level. Following the same procedure, we also explored other potential influencing factors of PE in separate linear regression models, including age group (60–69 years vs. 70–85 years), sex, education level (below/above upper secondary education), *APOE*-ε4 (carrier vs. non-carrier), test-retest time interval (1-3 months vs. 4-6 months), MRI parameters (below/above mean), National Adult Reading Test score (below/above median), and initial RBANS scores (below/above mean).

To assess the robustness of our main findings, we conducted the following sensitivity analyses: (1) modelling MRI parameters, age, test-retest time interval, initial RBANS score and NART score as continuous variables instead of dichotomised variables when exploring their associations with PE; (2) excluding 52 participants who waited for over 6 months after the screening visit to attend the baseline visit to avoid the loss of PE or occurrence of possible cognitive decline during the prolonged time gap; (3) additionally adjusting for test-retest time interval and modality of amyloid (PET or CSF) when assessing the amyloid-PE association.

Finally, to explore the predictive value of PE, PE was also assessed as a predictor together with initial RBANS score for amyloid positive status using binary logistic regression, adjusting for age, sex, education level, and *APOE*-ε4 carriage. The odds ratio (OR) and 95% confidence interval (CI) of standardised PE scores (i.e., centred and scaled) was reported, which reflects the relative risk of the presence of amyloid pathology per 1 SD increase in PE.

Statistical analyses were conducted using Stata (version 15; College Station, TX: StataCorp LLC). All statistical analyses are two-sided. A P value of < 0.05 indicates a statistically significant result.

RESULTS

Population Characteristics

Of the 502 participants assessed with RBANS scale in both screening and baseline clinic visits with median time gap of 3.5 months (interquartile range: 2.9–4.4), the mean (SD) age was 71.4 (5.5) years, and 254 (50.6%) were females. 192 participants (38.2%) were *APOE*-ε4 carriers and 247 (49.2%) were Aβ positive based on CSF Aβ42 level or PET scans. Nearly all participants (95.8%) were White. Most participants (85.7%) had completed upper secondary education or above.

Participant characteristics are presented by amyloid pathology status in **Table 1**. Aβ+ participants were slightly older and more likely to be *APOE*-ε4 carriers compared with Aβ- participants ($P < 0.05$). Differences in MRI parameters were also observed between amyloid groups, with Aβ+ group having lower hippocampal volume, whole brain volume, and AD signature cortical thickness ($P < 0.05$). The RBANS test-retest time interval was similar between Aβ+ group and Aβ- group ($P = 0.728$).

Practice Effects in Repeatable Battery for the Assessment of Neuropsychological Status Assessment

The internal consistency reliability of RBANS scale in our study sample measured by Cronbach's α was 0.64, and the test-retest reliability measured by Pearson correlation coefficient *r* was 0.79.

TABLE 1 | Population characteristics by amyloid status (*N* = 502).

Characteristics	Total	Amyloid positive	Amyloid negative	*P*-value
N	502	247	255	
Age (years), $\bar{x} \pm$ SD	71.4 ± 5.5	72.3 ± 5.6	70.4 ± 5.4	< 0.001
Female, %	50.6	48.6	52.6	0.374
Ethnicity (White), %	95.8	96.8	94.9	0.298
Below upper secondary education, %	14.3	17.0	11.8	0.094
APOE-ε4 carrier, %	38.2	54.7	22.4	< 0.001
NART score, $\bar{x} \pm$ SD	9.9 ± 6.7	9.5 ± 5.9	10.3 ± 7.3	0.202
Days between test and retest, median (IQR)	107 (87–133)	106 (86–133)	108 (87–136)	0.728
RBANS score (first test), $\bar{x} \pm$ SD				
Total scale	102.7 ± 11.8	102.6 ± 11.7	102.9 ± 11.9	0.734
Immediate memory index	101.6 ± 12.7	101.0 ± 12.3	102.2 ± 13.0	0.268
Delayed memory index	100.7 ± 10.1	99.8 ± 10.9	101.6 ± 9.2	0.045
Visuospatial construction index	95.7 ± 14.1	96.7 ± 13.9	94.9 ± 14.3	0.148
Language index	104.1 ± 11.5	104.5 ± 11.0	103.7 ± 12.0	0.422
Attention index	108.8 ± 14.5	108.5 ± 13.8	109.2 ± 15.2	0.607
MRI parameters, $\bar{x} \pm$ SD				
Hippocampal volume (mm^3)	7754 ± 852	7621 ± 899	7883 ± 794	< 0.001
Whole brain volume (mL^3)	1094629 ± 107552	1087603 ± 109462	1101408 ± 105861	0.005
Ventricular volume (mL^3)	35701 ± 16987	36381 ± 16991	35045 ± 16886	0.304
AD signature cortical thickness (mm)	2.80 ± 0.12	2.79 ± 0.13	2.81 ± 0.12	0.028

SD, standard deviation; NART, National Adult Reading Test; IQR, interquartile range; RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; MRI, magnetic resonance imaging; AD, Alzheimer's disease. P-values were calculated by chi-squared tests, t-tests, rank-sum test, or general linear regressions to adjust for intracranial volume for volumetric MRI parameters.

Participants had significantly higher scores in RBANS total scale and immediate and delayed memory indices in the second test than in the initial test (increased score = 3.9, 7.6, and 3.3, respectively; $P < 0.001$; **Table 2**). After taking into account the differences in variances of these indices, the calculation of within-subject Cohen's d_z revealed a strong effect size for PE in immediate memory index (0.70), and a low-to-moderate effect size for PE in RBANS total scale (0.48) and delayed memory index (0.35). In contrast, no significant PEs were identified for the rest of the three domain indices (Cohen's d_z ranged from 0.05 to 0.06; $P > 0.05$; **Table 2**).

Practice Effects in Repeatable Battery for the Assessment of Neuropsychological Status by Amyloid Pathology Status

We examined the practice effects in RBANS total scale and memory indices by amyloid pathology status (**Figure 1**). After adjusting for potential confounding factors, the amyloid positive group had significantly lower PE in immediate memory index than the amyloid negative group (Cohen's d_z = 0.60 vs. 0.81; $P = 0.022$). Similarly, a borderline statistical significance was observed for lower PE in delayed memory index, in the amyloid positive group (Cohen's d_z = 0.26 vs. 0.44; $P = 0.059$). However, the difference in PE in RBANS total scale by amyloid status did not reach statistical significance (Cohen's d_z = 0.46 vs. 0.50; $P = 0.387$; **Figure 1**). We also generated spaghetti plots by amyloid status to visualise the heterogeneity in practice effects across individuals (**Supplementary Figures 1–3**).

Other Influencing Factors on Practice Effect in Repeatable Battery for the Assessment of Neuropsychological Status

In the exploratory analyses for brain MRI parameters and PE, we observed no significant associations of hippocampal volume, whole brain volume, ventricular volume or AD signature cortical thickness with the magnitude of PE in RBANS total scale or memory indices (**Supplementary Table 1**).

Older adults (≥70 years), women, and *APOE*-ε4 non-carriers had larger PE in one or more RBANS indices ($P < 0.05$; **Table 3**). Those with worse performance in the initial RBANS test had larger PE in both total scale and the individual memory indices ($P < 0.05$; **Table 3**). Test-retest time interval, education level and NART score had no significant association with the magnitude of PE (**Supplementary Table 1**). Sensitivity analyses revealed consistent results with the main findings (**Supplementary Tables 2–5**).

Attenuated Practice Effect Is Indicative of Above Threshold Amyloid Pathology

We further explored the indicative value of PE for brain amyloid pathology. Results of multiple logistic regressions showed that, besides age (OR = 1.09, 95% CI: 1.05–1.13 per year) and *APOE*-ε4 carriage (OR = 5.50, 95% CI: 3.60–8.40), worse initial performance and lower PE in delayed memory index were independent predictors for amyloid positivity, with similar magnitudes of association (OR per 1

TABLE 2 | Differences between test and retest performance in repeatable battery for the assessment of neuropsychological status (RBANS) (N = 502).

RBANS score, $\bar{x} \pm$ SD	Test	Retest	Difference score (mean)	Difference score (range)	Cohen's d_z	*P*-value
Total scale	102.7 ± 11.8	106.6 ± 12.9	3.9	–20, 38	0.48	< 0.001
Immediate memory index	101.6 ± 12.7	109.2 ± 13.4	7.6	–28, 35	0.70	< 0.001
Delayed memory index	100.7 ± 10.1	104.0 ± 10.6	3.3	–35, 36	0.35	< 0.001
Visuospatial construction index	95.8 ± 14.1	96.6 ± 14.3	0.8	–37, 41	0.06	0.176
Language index	104.1 ± 11.5	104.8 ± 13.0	0.7	–42, 41	0.06	0.209
Attention index	108.8 ± 14.5	109.3 ± 14.7	0.5	–31, 32	0.05	0.293

RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; SD, standard deviation. P-values were calculated by paired t-tests.

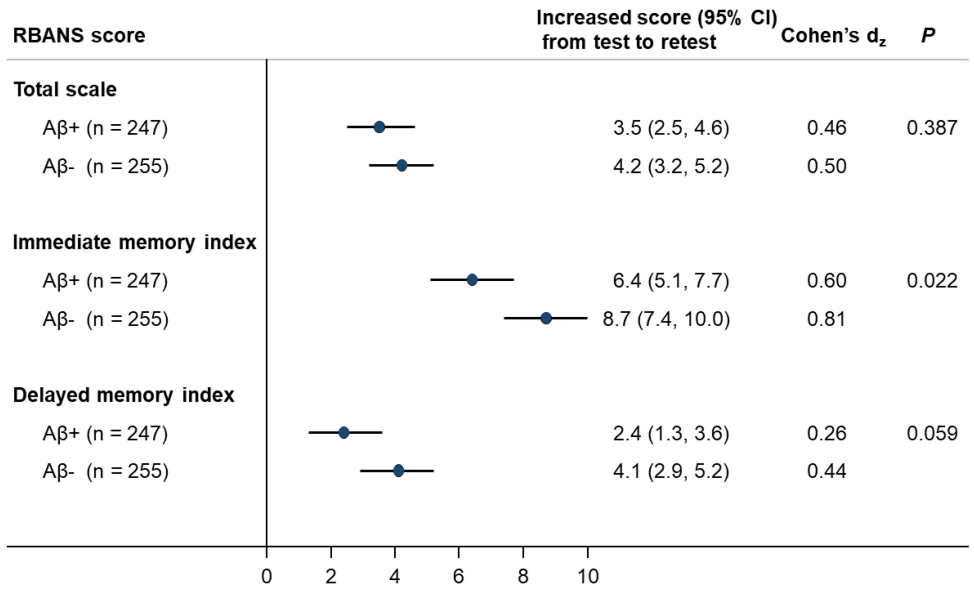


FIGURE 1 | Associations between amyloid status with magnitude of RBANS practice effects (N = 502). RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; CI, confidence interval. Estimates were adjusted for age, sex, education level, *APOE*-ε4 carriage and initial RBANS level, where applicable.

SD increase = 0.78, 95% CI: 0.63-0.97). As for immediate memory, lower PE (OR = 0.75, 95% CI: 0.61–0.94) but not performance in the initial test (OR = 0.82, 95% CI: 0.66–1.02) was a significant predictor for amyloid positivity. We did not find an association between PE in RBANS total scale and existing amyloid pathology (OR = 0.92, 95% CI: 0.75–1.12).

DISCUSSION

In this prospective cohort study of cognitively unimpaired older adults, enriched with fluid and neuroimaging biomarker data, we comprehensively assessed the practice effect in RBANS assessment and its potential influencing factors, with a focus on brain amyloid pathology. We observed significant practice effects for RBANS total scale and two memory indices, where participants performed better after repeated measurement using alternate versions of these tasks. The magnitude of practice effects differed by amyloid pathology status, age, sex, *APOE*-ε4 carriage and initial RBANS scores, but had no association with brain MRI parameters, education level or NART score.

Our findings suggest that PE in cognitive tests may be domain-specific. Of the five cognitive domains assessed by RBANS scale, only the two memory indices presented significant practice effects, whilst participants' performance in visuospatial construction, language and attention domains remained similar between the first and second tests over a median of 3.5 months. Our results were in line with a previous study of a much smaller sample of 36 healthy adults (Bartels et al., 2010), where clinically relevant PE was observed during high-frequency testing within three months in learning and memory tests but not in language and visuospatial tests. Similarly, a study of 947 cognitively normal older adults from the Mayo Clinic Study of Aging showed large PE in learning and memory tests but low PE in language tests, using the Mayo Clinic neurocognitive battery (Machulda et al., 2013).

Regarding the memory domain indices, we observed a much larger effect size of PE for immediate memory index than that for delayed memory index or the RBANS total scale. This implies

TABLE 3 | Associations between other characteristics and magnitude of repeatable battery for the assessment of neuropsychological status (RBANS) practice effects (*N* = 502).

Characteristics	No. of participants	Increase of total scale (95% CI)	*P*-value	Increase of immediate memory index (95% CI)	*P*-value	Increase of delayed memory index (95% CI)	*P*-value
Age (years)			0.565		0.146		0.009
60–69	208	4.1 (3.0, 5.2)		6.8 (5.3, 8.2)		2.0 (0.8, 3.2)	
70–85	294	3.7 (2.8, 4.6)		8.1 (6.9, 9.3)		4.2 (3.1, 5.2)	
Sex			0.018		0.002		0.712
Male	248	3.0 (2.0, 4.0)		6.1 (4.8, 7.4)		3.1 (2.0, 4.2)	
Female	254	4.7 (3.7, 5.7)		9.0 (7.7, 10.2)		3.4 (2.3, 4.6)	
APOE-ε4			0.164		0.004		0.337
Carrier	192	3.2 (2.1, 4.4)		5.9 (4.4, 7.3)		3.8 (2.5, 5.0)	
Non-carrier	310	4.3 (3.4, 5.2)		8.6 (7.5, 9.8)		3.0 (2.0, 4.0)	
Initial RBANS score			0.002		< 0.001		< 0.001
Higher than mean level	238	2.7 (1.6, 3.7)		4.7 (3.5, 6.0)		0.9 (–0.2, 1.9)	
Lower than mean level	264	5.0 (4.0, 5.9)		10.7 (9.3, 12.0)		6.2 (5.0, 7.4)	

RBANS, Repeatable Battery for the Assessment of Neuropsychological Status; CI, confidence interval. Estimates were adjusted for age, sex, education level, APOE-ε4 carriage and initial RBANS level, where applicable.

that PE may be more pronounced in immediate memory tasks where people tend to get better at doing these tasks following familiarisation with the test materials or procedures, even when assessed with different word lists (Houx et al., 2002). Thus, the immediate memory test seems to be a more sensitive measure of PE, compared with other domains or the global composite score. The contrast between immediate and delayed memory PEs might alternatively reflect differences in the content of the measures. Specifically, the RBANS immediate memory index is derived solely from tests of verbal recall, whereas the delayed memory index also incorporates verbal recognition and visual-constructional recall. Future systematic evaluation of practice effects in individual test scores rather than the overall indices, with larger sample size and careful control of multiple testing, may help identify even more sensitive metrics.

Our data are in line with previous reports, suggesting the predictive value of PE for the presence of amyloid pathology and subsequent cognitive decline, in addition to merely evaluating cognitive measurement. To be noted, on average, the RBANS scores in our study participants were within "cognitively healthy" boundaries, even in the amyloid positive group and would not prompt further testing in a clinical scenario. This observation underscores the potential value of diminished practice effects as an adjunct metric to traditional assessments for the sensitive detection of preclinical AD. Several previous studies have consistently shown that diminished PE over repeated cognitive testing (mainly episodic memory measures) was associated with subsequent cognitive decline and increased risk of mild cognitive impairment (MCI) or dementia (Duff et al., 2007; Sanchez-Benavides et al., 2016; Jutten et al., 2020, 2021). In contrast, previous evidence on the association between PE and AD biomarkers and neuropathology remained inconsistent (Duff et al., 2018; Ihara et al., 2018; Jutten et al., 2020). A previous systematic review on PE in cognitive assessment identified four papers reporting an association between higher amyloid uptake on amyloid PET scans and lower PE, whereas two papers did not detect this association (Jutten et al., 2020). In our study, the attenuated PE in memory indices was associated with the presence of high amyloid burden but not with brain MRI features, including hippocampal volume, implying that PE in memory tests could be more indicative of β-amyloidosis [which is specific for Alzheimer's disease (AD)] instead of biomarkers of neurodegeneration or neuronal injury (Jack et al., 2016). Consistent with our results, a recent report from the Harvard Aging Brain Study, of 114 cognitively unimpaired older adults, showed that lower PE in a self-administered computerised cognitive composite battery over the first 3 months was associated with more global amyloid burden (based on PiB-PET imaging) and tau deposition in the entorhinal cortex and inferior-temporal lobe (based on Flortaucipir PET imaging) (Jutten et al., 2021). These findings imply the usefulness of PE as an early detection tool for signs of disease burden prior to the emergence of cognitive impairment, which might inform participant stratification and biomarker testing strategies for clinical trials.

In our exploratory analyses, practice effects in RBANS total scale or memory indices were more pronounced in older adults, women, *APOE*-ε4 non-carriers and those with worse performance in the initial RBANS assessment (probably due to larger room for improvement). Of note, these factors were associated with different indices, indicating a complex domain-specific PE population heterogeneity. Our finding of a positive association between age and PE was inconsistent with a previous meta-analysis report (Calamia et al., 2012) of a negative association, in a much younger population (mean age of around 40 to 50 years). In the afore-mentioned Mayo Clinic report (Machulda et al., 2013), no significant PE differences were found on memory test scores between those aged below and above 80 years. A previous systematic review identified three papers reporting an association between presence of ≥ 1 *APOE*-ε4 allele and lower PE, whereas three papers did not detect this association (Jutten et al., 2020). Further studies are warranted to elucidate the nature and extent of these population heterogeneities in PE, which could be crucial for clinical trials in obtaining unbiased

effect estimate for tested treatment or intervention. If the factors affecting PE are not well balanced between placebo and treatment groups, the two groups may have different levels of PE, in which case researchers need to control for these factors so that the estimate of difference in cognitive outcomes between groups can be attributed to treatment.

The availability of extensive phenotypic (including fluid and neuroimaging biomarker) data is a key strength of our study. Moreover, the relatively short test-retest interval (median of 3.5 months) was essential in minimising the risk of a potential cognitive decline during the test-retest interval affecting the presence and extent of PE. If given a long test-retest period, PE may be masked by progressive cognitive decline over time and it would be difficult to distinguish one from the other.

Several limitations need to be taken into consideration when interpreting our results. Since we explored multiple influencing factors on PE in our study, the risk of inflated Type 1 error in multiple testing cannot be ruled out. Therefore, our exploratory analyses need further validation. Moreover, RBANS does not provide an isolated scale of executive function, a domain which has been independently associated with early amyloidosis rather than memory performance decrements in cognitively normal adults (Tideman et al., 2022). Assessing diminished practice effects in this domain may yet provide even more sensitive markers of subtle cognitive signs. Due to the different modalities and tracers used for amyloid testing in this study, we did not evaluate the amyloid pathology on a quantitative scale which is worth to be considered in future studies. In addition, we only used data from two time points; future studies on longitudinal PE across multiple measurements (with short between-test intervals) are needed. For instance, it is worth exploring whether the PE beyond the second test is not as large as that between the first two tests, which may have important implications for research and clinical purposes (e.g., recommending the second assessment to be considered as baseline measure to minimise PE in outcome assessment). Furthermore, since our test-retest time gap mainly fell between 3 and 4 months, future large-scale studies with time gaps of wider distribution could provide insights for what might be too short vs. too long for detecting PE, though it is possible that the optimal time gap could be different for different cognitive domains or tasks. Finally, our study population are cognitively unimpaired older adults; it would also be interesting to investigate PE in MCI or AD patients, which may show different profiles (Machulda et al., 2013). Similarly, the study sample lacks ethnic and racial diversity (95.8% White people) thereby limiting the generalisability of our findings.

In conclusion, we identified significant PE in RBANS total scale and memory indices among a cohort of cognitively unimpaired older adults. PE is not simply a source of measurement bias in cognitive assessment, but may be informative with regard to a significant brain amyloid pathology burden.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available at present due to embargo on the data. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

This study received approval from the National Research Ethics Service (NRES) Committee London Central (reference 15/LO/0711 (IRAS 140764)), as well as independent ethics review by committees from the local sites. All participants provided informed consent before participating in the study.

AUTHOR CONTRIBUTIONS

GP, LM, CU-M, BZ, and TW contributed to study design and conception. BZ and CU-M carried out data analysis and interpretation. BZ, LM, and CU-M drafted the first version of the manuscript. All authors critically reviewed and revised the manuscript.

FUNDING

This study was supported by Janssen Research & Development, United States.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.909614/full#supplementary-material

REFERENCES

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., and Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neurosci.* 11:118. doi: 10.1186/1471-2202-11-118

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Duff, K., Anderson, J. S., Mallik, A. K., Suhrie, K. R., Atkinson, T. J., Dalley, B. C. A., et al. (2018). Short-term repeat cognitive testing and its relationship to hippocampal volumes in older adults. *J. Clin. Neurosci.* 57, 121–125. doi: 10.1016/j.jocn.2018.08.015

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., Haase, R. F., et al. (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch. Clin. Neuropsychol.* 22, 15–24. doi: 10.1016/j.acn.2006.08.013

Elias, M. F., Beiser, A., Wolf, P. A., Au, R., White, R. F., and D'Agostino, R. B. (2000). The preclinical phase of alzheimer disease: a 22-year prospective study of the Framingham cohort. *Arch. Neurol.* 57, 808–813. doi: 10.1001/archneur.57.6.808

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications

for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement.* 1, 103–111. doi: 10.1016/j.dadm.2014.11.003

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., and Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *J. Appl. Psychol.* 92, 373–385. doi: 10.1037/0021-9010.92.2.373

Houx, P. J., Shepherd, J., Blauw, G. J., Murphy, M. B., Ford, I., Bollen, E. L., et al. (2002). Testing cognitive function in elderly populations: the PROSPER study. Prospective study of pravastatin in the elderly at risk. *J. Neurol. Neurosurg. Psychiatry* 73, 385–389. doi: 10.1136/jnnp.73.4.385

Ihara, R., Iwata, A., Suzuki, K., Ikeuchi, T., Kuwano, R., Iwatsubo, T., et al. (2018). Clinical and cognitive characteristics of preclinical Alzheimer's disease in the Japanese Alzheimer's disease neuroimaging initiative cohort. *Alzheimers Dement.* 4, 645–651. doi: 10.1016/j.trci.2018.10.004

Jack, C. R. Jr, Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., et al. (2016). A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* 87, 539–547. doi: 10.1212/WNL.0000000000002923

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S. E., et al. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimers Dement.* 12:e12055. doi: 10.1002/dad2.12055

Jutten, R. J., Rentz, D. M., Fu, J. F., Mayblyum, D. V., Amariglio, R. E., Buckley, R. F., et al. (2021). Monthly at-home computerized cognitive testing to detect diminished practice effects in preclinical Alzheimer's disease. *Front. Aging Neurosci.* 13:800126. doi: 10.3389/fnagi.2021.800126

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

Nalder, L., Zheng, B., Chiandet, G., Middleton, L. T., and de Jager, C. A. (2021). Vitamin B12 and folate status in cognitively healthy older adults and associations with cognitive performance. *J. Nutr. Health Aging* 25, 287–294. doi: 10.1007/s12603-020-1489-y

Nelson, H. E., and Willison, J. (1991). *National Adult Reading Test (NART)*. Windsor: Nfer-Nelson.

Randolph, C. (1998). *Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)*. San Antonio, TX: Psychological Corporation.

Randolph, C., Tierney, M. C., Mohr, E., and Chase, T. N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): preliminary clinical validity. *J. Clin. Exp. Neuropsychol.* 20, 310–319. doi: 10.1076/jcen.20.3.310.823

Sanchez-Benavides, G., Gispert, J. D., Fauria, K., Molinuevo, J. L., and Gramunt, N. (2016). Modeling practice effects in healthy middle-aged participants of the Alzheimer and families parent cohort. *Alzheimers Dement.* 4, 149–158. doi: 10.1016/j.dadm.2016.07.001

Sanderson-Cimino, M., Elman, J. A., and Tu, X. (2022). Practice effects in mild cognitive impairment increase reversion rates and delay detection of new impairments. *Front. Aging Neurosci.* 14:847315. doi: 10.3389/fnagi.2022.847315

Schwarz, C. G., Gunter, J. L., Wiste, H. J., Przybelski, S. A., Weigand, S. D., Ward, C. P., et al. (2016). A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. *Neuroimage Clin.* 11, 802–812. doi: 10.1016/j.nicl.2016.05.017

Tideman, P., Stomrud, E., Leuzy, A., Mattsson-Carlgren, N., Palmqvist, S., Hansson, O., et al. (2022). Association of β-Amyloid accumulation with executive function in adults with unimpaired cognition. *Neurology* 98, e1525–e1533. doi: 10.1212/WNL.0000000000013299

Udeh-Momoh, C. T., Watermeyer, T., Price, G., de Jager Loots, C. A., Reglinska-Matveyev, N., Ropacki, M., et al. (2021). Protocol of the cognitive health in ageing register: investigational, observational and trial studies in dementia research (CHARIOT): prospective readiness cohort (PRO) substudy. *BMJ Open* 11:e043114. doi: 10.1136/bmjopen-2020-043114

Conflict of Interest: SB and GN were employed by Janssen Research and Development LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from Janssen Research & Development, USA. This is a collaborative study with the sponsor's clinicians and scientists.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Aging Neuroscience

BRIEF RESEARCH REPORT
published: 18 July 2022
doi: 10.3389/fnagi.2022.838459

Neuropsychological Decline Stratifies Dementia Risk in Cognitively Unimpaired and Impaired Older Adults

Jean K. Ho[1] *and Daniel A. Nation*[1,2]*

[1] *Institute for Memory Disorders and Neurological Impairments, University of California, Irvine, Irvine, CA, United States,*
[2] *Department of Psychological Science, University of California, Irvine, Irvine, CA, United States*

Objective: Validation and widespread use of markers indicating decline in serial neuropsychological exams has remained elusive despite potential value in prognostic and treatment decision-making. This study aimed to operationalize neuropsychological decline, termed "neuropsychological (NP) decline," in older adults followed over 12 months in order to aid in the stratification of dementia risk along the cognitively unimpaired-to-mild cognitive impairment (MCI) spectrum.

Methods: A prospective cohort study utilized 6,794 older adults from the National Alzheimer's Coordinating Center (NACC) database with a baseline diagnosis of normal cognition, impaired without MCI or with MCI. Operationalization of NP decline over 12-month follow-up used regression-based norms developed in a robustly normal reference sample. The extent to which each participant's 12-month follow-up score deviated from norm-referenced expectations was quantified and standardized to an NP decline z-score. Cox regression evaluated whether the NP decline metric predicted future dementia.

Results: Participant's NP decline scores predicted future all-cause dementia in the total sample, $\chi^2 = 110.71$, hazard ratio (HR) = 1.989, $p < 0.001$, and in the subset diagnosed with normal cognition, $\chi^2 = 40.84$, HR = 2.006, $p < 0.001$, impaired without MCI diagnosis, $\chi^2 = 14.89$, HR = 2.465, $p < 0.001$, and impaired with MCI diagnosis, $\chi^2 = 55.78$, HR = 1.916, $p < 0.001$.

Conclusion: Operationalizing NP decline over 12 months with a regression-based norming method allows for further stratification of dementia risk along the cognitively unimpaired-to-MCI spectrum. The use of NP decline as an adjunctive marker of risk beyond standard cognitive diagnostic practices may aid in prognosis and clinical decision-making.

Keywords: mild cognitive impairment, subtle cognitive decline, dementia, Alzheimer's disease, aging, assessment

OPEN ACCESS

Edited by:
Joel Ramirez,
University of Toronto, Canada

Reviewed by:
Julie Suhr,
Ohio University, United States
Kevin Duff,
The University of Utah, United States

***Correspondence:**
Daniel A. Nation
dnation@uci.edu

Specialty section:
This article was submitted to Alzheimer's Disease and Related Dementias, a section of the journal Frontiers in Aging Neuroscience

Received: *17 December 2021*
Accepted: *20 June 2022*
Published: *18 July 2022*

Citation:
Ho JK and Nation DA (2022) Neuropsychological Decline Stratifies Dementia Risk in Cognitively Unimpaired and Impaired Older Adults. Front. Aging Neurosci. 14:838459. doi: 10.3389/fnagi.2022.838459

INTRODUCTION

Early identification of older adults at risk for dementia remains an important research goal, as preventative efforts will likely require early intervention (Crous-Bou et al., 2017). Although mild cognitive impairment (MCI) is an important and useful diagnostic construct that represents an intermediate level of cognitive impairment between normal cognition and dementia (Petersen, 2011), recent research has increasingly focused on earlier stratification of dementia risk in cognitively unimpaired older adults (Amieva et al., 2005; Machulda et al., 2013; Hassenstab et al., 2015; Han et al., 2017). Efforts aimed at identifying cognitively unimpaired older adults at risk for dementia have predominantly emphasized the role of biological markers in index underlying neuropathology (Jack et al., 2018). However, numerous studies also indicate that subtle cognitive changes are detectable on a neuropsychological exam in cognitively unimpaired older adults at risk for dementia (Edmonds et al., 2015b; Han et al., 2017; Ho and Nation, 2018; Thomas et al., 2020).

There are inherent limits in the ability to establish cutoff values and diagnostic criteria for the diagnosis of subtle or mild impairments based on a single exam. Thus, longitudinal assessment of cognitive change within an individual may aid in the detection of early decline within normal range performance (Koscik et al., 2019; Nation et al., 2019). However, serial cognitive exams introduce practice effects and regression to the mean, complicating the interpretation of decline (Crawford and Howell, 1998; Slick, 2006). Nevertheless, recent studies suggest that serial cognitive performance may still be of value. For example, the lack of a practice effect may actually be indicative of a subtle cognitive decline in older adults at risk for dementia (Machulda et al., 2013; Hassenstab et al., 2015; Duff et al., 2017; Papp et al., 2020). These findings suggest the potential value of obtaining normative data on serial cognitive exam performance in older adults to supplement single exam data.

Obtaining information regarding the trajectory of cognitive change may aid efforts to refine MCI diagnostic accuracy and predictive value (Nation et al., 2019). Fluctuation in cognitive performance and reversion from MCI to normal performance across exams is common, even among individuals with underlying neuropathology (Thomas et al., 2019). If the trajectory of cognitive change was available in patients with MCI through normative comparisons of cognitive change, further characterization of MCI-associated risk could be possible.

To evaluate the predictive value of serial neuropsychological exam analysis, we previously operationalized neuropsychological decline, termed "NP decline," over 1 year using the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (Nation et al., 2019). In this study, NP decline in cognitively unimpaired older adults, and those diagnosed with MCI, was associated with an increased risk for future clinical diagnosis of Alzheimer's dementia. This study sought to further validate this previously developed NP decline metric and determine its predictive value for all-cause dementia. We hypothesized that, consistent with our previous results, NP decline would be predictive of future Alzheimer's disease, even in a larger and more heterogeneous sample of 6,794 older adults from the National Alzheimer's Coordinating Center (NACC) database.

METHODS

National Alzheimer's Coordinating Center Study Data and Participants

This prospective cohort study utilized longitudinal participant data obtained from the NACC database, a repository of data on aging and dementia gathered from Alzheimer's Disease Centers (ADCs) across the country using a Uniform Data Set (UDS). The UDS includes harmonized protocols for data collection and entry regarding information from in-person visits for health and neurological examination, neuropsychological testing, and psychosocial and biological measures. In this study, NACC UDS data from the cognitive diagnostic exam and neuropsychological exam were analyzed, and all available follow-up data through December 2018 were included. The duration of available participant follow-up data varied from 18 to 156 months after baseline. Given the switch in verbal memory measures between UDS 2.0 and 3.0, we included data from Logical Memory only and did not include Craft Story data.

We limited our analysis to the 6,794 participants who were aged 60 years and older, had been diagnosed "cognitively normal," "impaired without MCI" or "MCI," according to the NACC UDS protocol criteria, and had been followed for at least two additional follow-up study visits extending more than 12 months from baseline. All participants needed 12-month follow-up data in order to calculate NP decline scores and needed to remain non-demented at a 12-month follow-up in order to be included in the analysis of 12-month NP decline as a predictor of future dementia. Similarly, all participants required the third evaluation after their 12-month follow-up exam in order to be evaluated in terms of the predictive value of a 12-month NP decline for the risk for future dementia. Thus, participants who progressed to dementia within 12 months of baseline, had fewer than 3 exams, or had less than $\leq$12 months of total follow-up were excluded.

Participants from NACC are assigned a diagnosis following adjudication by an experienced clinician or an interdisciplinary team (Morris et al., 2006). Psychosocial functioning, history, as well as test performance in various cognitive domains (recall, attention, executive function, language, and visuospatial functioning) are under consideration during these adjudications. Diagnoses in NACC are informed by neuropsychological testing, but are made clinically and are not based on strict cutoff values on these measures. Participants receive a diagnosis of (a) "cognitively normal" if they lack significant functional or cognitive impairment, (b) "MCI" if they have subjective or objective evidence of cognitive impairment without significant functional impairment, and (c) "demented" if they have both significant functional and cognitive impairment.

All contributing ADCs obtained informed consent from their participants and maintained separate IRB review and approval from their institutions prior to submitting data to NACC. Recruitment methods and sample characteristics varied across each ADC, representing a mixture of clinical- and community-based sampling.

Baseline Versus 12-Month Diagnoses

For all analyses, participant clinical diagnostic groups were determined based on the 12-month follow-up examination to ensure that NP decline fell within the range of the appropriate diagnostic classifications (i.e., decline within normal range cognition, decline within no MCI range cognitive impairment, and decline within MCI range cognition).

Regression-Based Norms for Neuropsychological Decline Using the Alzheimer's Disease Neuroimaging Initiative Database

To avoid circularity in our investigation into the predictive utility of a neuropsychological marker for future dementia risk (i.e., NP decline), we first developed the NP decline marker using normative data from a reference sample in one dataset (the ADNI data) and then applied these norms to a separate test sample from another dataset (the NACC data). To avoid circularity and criterion contamination of clinical diagnosis by the neuropsychological markers themselves, all findings were also confirmed using progression from a CDR® Dementia Staging Instrument score of 0 to a score of 0.5 or higher as the criterion measure, rather than clinical diagnosis.

The NP decline metric was operationalized by developing linear regression equations in a robustly normal reference sample from the ADNI database (n = 294). For this analysis, we used methods described in detail recently (Nation et al., 2019). Briefly, a robustly normal subset of cognitively normal older adults from the ADNI study was identified using criteria established by prior ADNI studies (Edmonds et al., 2015a): (1) participants were identified as cognitively normal on baseline ADNI assessment and (2) participants remained cognitively normal throughout the duration of their study participation.

Linear regression was used to model the relationship between baseline performance on a neuropsychological test and 12-month follow-up performance on the same test using longitudinal ADNI study data. Neuropsychological tests included Wechsler Memory Scale – Revised (WMS-R) Logical Memory Story A immediate (Logical Memory I) and delayed (Logical Memory II) free recall, Trails A and B, and Animals and Vegetables. Specific neuropsychological tests were chosen based on the overlap between ADNI neuropsychological tests (reference sample) and tests available in NACC (test sample), as well as the desire to evaluate a balance of 2 tests per domain across domains relevant to dementia risk, including memory, attention/executive function, and language (Bondi et al., 2008). Scores from Trails A and B exhibited significant skewness, which was corrected by log transformation. These scores were also inverted (i.e., multiplied by -1) such that higher scores indicate better performance, consistent with all other neuropsychological measures.

The result of linear regression analyses evaluating baseline test performance as a predictor of 12-month follow-up test performance produced linear regression equations that represent the relationship between baseline and 12-month test performance in a robustly normal sample (refer to **Supplementary Table 1** for details regarding linear regression parameters in the robustly normal ADNI sample). These regression-based norms were developed for the purpose of calculating standardized scores for NP decline over 12 months relative to normative expectations (as in Nation et al., 2019). This study sought to apply these ADNI-derived regression-based norms to a test sample from the NACC database to determine whether the resulting NP decline metric may be of value in predicting future dementia among older adults who were cognitively normal or mildly impaired during their first 12 months of neuropsychological follow-up.

Applying Regression-Based Norms From Alzheimer's Disease Neuroimaging Initiative to the National Alzheimer's Coordinating Center Database

In this study, the linear regression equations developed in the robustly normal sample from ADNI (refer to earlier) were used to quantify NP decline scores for all eligible participants in the NACC database with a baseline clinical consensus diagnosis of normal cognition, impaired without MCI or MCI. Below, Eq. 1 shows the template for the normative regression equations developed from raw scores in robustly normal participants in ADNI and used to calculate the predicted 12-month performance for each test for NACC participants (Eqs 2–7).

The NP decline metric was calculated as previously described using three steps (Nation et al., 2019), namely, (1) baseline NACC participant raw scores on neuropsychological testing (refer to earlier for battery) were entered into the linear regression equations (Eqs 2–7) developed using robustly normal participants from ADNI. Linear regression equations used baseline raw scores to calculate the predicted 12-month performance on each neuropsychological test based on normative expectations from the ADNI subsample. (2) For each participant, the predicted 12-month performance based on the regression-based norms from ADNI was then subtracted from the actual 12-month performance for each neuropsychological test, and the resulting discrepancy between the 12-month predicted performance and the actual performance was divided by the standard error of the estimate for each linear regression equation corresponding to each neuropsychological test (refer to Eq. 8 below). (3) The standardized scores were averaged across all 6 neuropsychological test scores to create the NP decline z-score.

As shown in Eq. 8, NP decline raw scores were standardized by dividing the standard error of the estimate ($S_{y.x}$) drawn from each regression equation (Crawford and Howell, 1998; Crawford and Garthwaite, 2006): Eq. 2 $S_{y.x}$ = 2.7730, Eq. 3 $S_{y.x}$ = 3.1780,

Eq. 4 $S_{y.x} = 0.1009$, Eq. 5 $S_{y.x} = 0.1374$, Eq. 6 $S_{y.x} = 4.0650$; and Eq. 7 $S_{y.x} = 3.2700$.

$$Predicted\ score = intercept + (coefficient \times baseline\ score) \quad (1)$$

$$Predicted\ Logical\ Memory\ I = 6.883 + (0.595 \times baseline\ Logical\ Memory\ I) \quad (2)$$

$$Predicted\ Logical\ Memory\ II = 4.810 + (0.680 \times baseline\ Logical\ Memory\ II) \quad (3)$$

$$Predicted\ Trails\ A\ log = [0.589 + (0.598 \times baseline\ Trails\ A\ log)] \times -1 \quad (4)$$

$$Predicted\ Trails\ B\ log = [0.656 + (0.643 \times baseline\ Trails\ B\ log)] \times -1 \quad (5)$$

$$Predicted\ Animals = 8.410 + (0.623 \times baseline\ Animals) \quad (6)$$

$$Predicted\ Vegetables = 4.464 + (0.687 \times baseline\ Vegetables) \quad (7)$$

$$NP\ decline\ subtest\ z = \frac{actual\ score - predicted\ score}{standard\ error\ of\ the\ estimate} \quad (8)$$

Individual Test Scores Versus Overall Neuropsychological Decline Score

The examination of NP decline in individual test scores is beyond the scope of this study, which is focused instead on NP decline as a general cognitive decline factor assessed by multiple test scores. The use of single test scores to determine clinical status is also not advised, given the limited reliability of individual neuropsychological test scores for determining cognitive abnormality (Binder et al., 2009). Finally, our prior study developed an optimized cutoff value for NP decline based on the overall average NP decline across tests (Nation et al., 2019), providing an opportunity for cross-validation using the NACC data. For all these reasons, NP decline subtest z-scores were averaged to create a global NP decline score for all statistical analyses, as described earlier.

Neuropsychological Decline Cutoff Values – Cross-Validation

The optimal cutoff values for NP decline in the ADNI study were previously determined by receiver operating characteristic (ROC) curve analysis. Results of the ROC curve analysis indicated an optimal NP decline z-score of -0.5808, corresponding approximately to the 28th percentile of the NP decline distribution (Nation et al., 2019). This z-score represents an optimal cutoff value for the NP decline metric in terms of predicting the development of dementia. It is a z-score of the distribution of NP decline, computed as predicted performance for normal aging subtracted from actual 12-month follow-up performance, and standardized by the standard error of the estimate. Cognitively normal older adults performing below this NP decline z-score at 12-month follow-up exhibited more rapid progression to dementia, relative to those above the cutoff value. This was regardless of demographic factors, biomarker status, or APOE4 carrier status (Nation et al., 2019). For cross-validation, this study used this same cutoff value derived from the ADNI study to determine dementia risk based on NP decline in the NACC sample.

Statistical Analyses

All study variables were evaluated for departures from normality and potentially influential outliers. Trails A and B scores were log-transformed to improve normality for the purposes of linear regression models of NP decline (refer to Eqs 4, 5 above).

Participants were divided into groups based on the combination of their 12-month NACC clinical diagnostic status (cognitively normal, impaired without MCI, and MCI) and their final diagnostic status (no dementia vs. dementia). Participant groups were compared on their baseline demographic and clinical measures, including age, sex, and education using a 2 × 2 (diagnostic status × NP decline status) ANCOVA controlling for age, sex, and years of education, with *post-hoc* Bonferroni-corrected pairwise comparisons. Chi-squared analyses were used to compare the rate of future dementia by clinical diagnostic and NP decline status. Cox regression was used to evaluate the predictive value of NP decline in the overall sample and within each clinical diagnostic group, controlling for age, sex, and education.

RESULTS

Participant demographics and clinical data are presented in **Table 1**. Cognitively normal older adults with greater than expected 12-month NP decline (below-established cutoff value) were significantly more likely to develop dementia over all follow-up relative to those above the cutoff value χ^2 (1, N = 4,692) = 55.02, $p < 0.00001$. Impaired without MCI participants with greater than expected 12-month NP decline (below-established cutoff value) were significantly more likely to ultimately develop dementia over all follow-up relative to those above the cutoff value χ^2 (1, N = 470) = 4.78, $p < 0.05$. Similarly, MCI participants with greater than expected 12-month NP decline (below-established cutoff value) were significantly more likely to ultimately develop dementia over all follow-up relative to those above the cutoff value χ^2 (1, N = 1,632) = 29.21, $p < 0.00001$.

Results of 2 × 2 ANCOVA (baseline clinical diagnosis × dementia outcome) with NP decline z-score as the dependent measure are presented in **Figure 1**. Cognitively normal older adults who ultimately developed dementia exhibited significantly worse NP decline than those who did not develop dementia ($p < 0.001$) and did not significantly differ

TABLE 1 | Participant demographics and clinical characteristics.

Demographics	Mean ± SD or n	Range or %	
Age (years)	74.01 ± 7.82	60-104	
Education (years)	15.52 ± 3.21	0-30	
Male to Female Ratio	2,618 to 4,176	38.5% male	
NACC Diagnosis at 12-months			
Normal Cognition	4,692	69.1%	
Impaired MCI−	470	6.9%	
Impaired MCI+	1,632	24.0%	
Progression to dementia			
Dementia at Follow up	764	11.2%	
Follow up (months)	58.97 ± 29.88	19-158	
NACC diagnosis × NP decline			**% Dementia conversion**
Normal/NP−	3,557	52.4%	4.2%
Normal/NP+	1,135	16.2%	10.0%
Impaired MCI−/NP−	308	4.5%	9.7%
Impaired MCI−/NP+	162	2.4%	16.7%
Impaired MCI+/NP−	738	10.9%	20.6%
Impaired MCI+/NP+	894	13.2%	32.6%

MCI−, Mild Cognitive Impairment absent; MCI+, Mild Cognitive Impairment present; NP−, Neuropsychological Decline absent; NP+, Neuropsychological Decline present; SD, standard deviation; NACC, National Alzheimer's Coordinating Center.

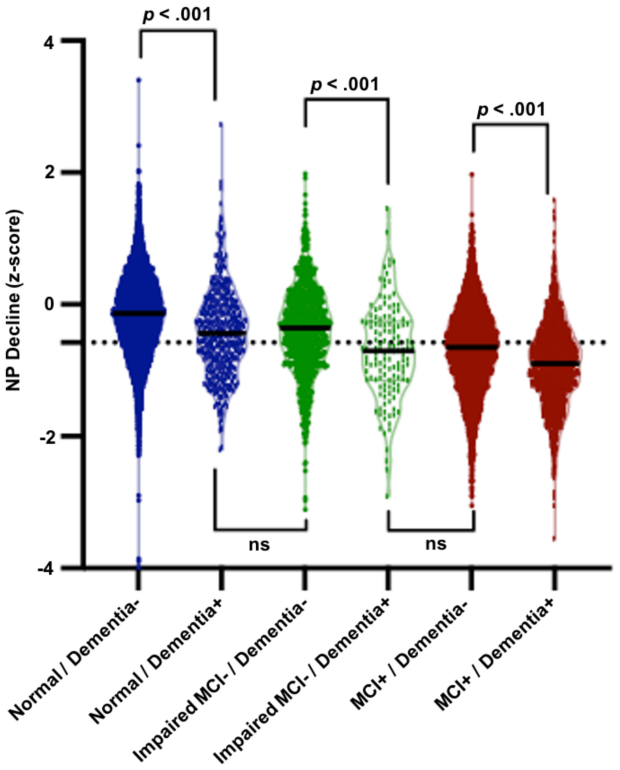


FIGURE 1 | Results of 2 × 2 ANCOVA (baseline clinical diagnosis × dementia outcome). Cognitively normal older adults who ultimately developed dementia exhibited significantly worse NP decline (M = –0.46, SD = 0.57, range = 3.32) than those who did not develop dementia (M = –0.14, SD = 0.59, range = 3.98, $p < 0.001$) and did not significantly differ in NP decline from those who were impaired without MCI. Similarly, impaired without MCI participants who progressed to dementia displayed worse NP decline (M = –0.58, SD = 0.60, range = 2.55) than those who did not progress to dementia (M = –0.31, SD = 0.64, range = 2.50, $p < 0.001$) and did not significantly differ from those who were diagnosed with MCI. Finally, MCI participants who progressed to dementia exhibited significantly greater NP decline (M = –0.79, SD = 0.58, range = 2.59) than those who did not progress to dementia (M = –0.58, SD = 0.63, range = 2.35, $p < 0.001$).

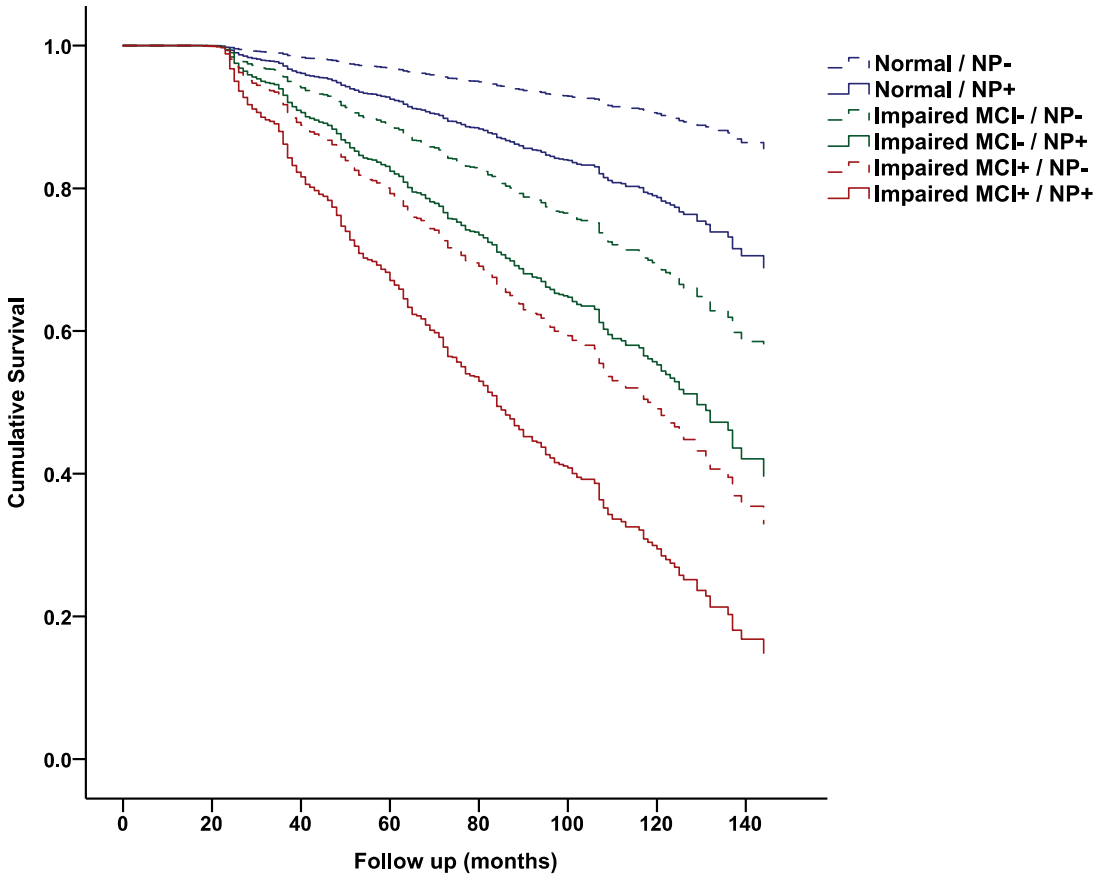
in NP decline from those who were impaired without MCI. Similarly, impaired without MCI participants who progressed to dementia displayed worse NP decline than those who did not progress to dementia ($p < 0.001$) and did not significantly differ from those who were diagnosed with MCI. Finally, MCI participants who progressed to dementia exhibited significantly greater NP decline than those who did not progress to dementia ($p < 0.001$).

On longitudinal analysis, NP decline predicted future all-cause dementia in the total sample, after controlling for age, sex, and education, -2 log likelihood = 11,874.363, χ^2 = 295.601.71, hazard ratio [HR] = 2.806, $p < 0.001$, and in the subset with normal cognition, -2 log likelihood = 3,776.938, χ^2 = 40.842, HR = 2.006, $p < 0.001$, impaired without MCI diagnosis, -2 log likelihood = 574.928, χ^2 = 14.891, HR = 2.465, $p < 0.001$, and impaired with MCI diagnosis, -2 log likelihood = 5,747.221, χ^2 = 55.772, HR = 1.916, $p < 0.001$. Results of Cox regression analysis stratified by clinical diagnosis and NP decline status are presented in **Figure 2**.

DISCUSSION

Among older adults with a baseline diagnosis spanning the cognitively unimpaired-to-MCI spectrum, NP decline indicative of worse than expected 12-month follow-up performance was associated with an approximately 2-fold increase in risk for all-cause dementia at each follow-up, even after accounting for age, sex, and education. Thus, NP decline may represent a valuable adjunctive tool for risk stratification in both normal and mildly impaired older adults followed for at least 12 months. Frequently used diagnostic criteria for MCI and for cognitive decline in the context of Alzheimer's disease rely heavily on subjective self-report and informant report to assess the presence of longitudinal decline (Jack et al., 2018), but subjective reports of cognitive change are influenced by psychiatric symptoms, personality traits, and other unrelated factors that may contribute to diagnostic error (Edmonds et al., 2014, Edmonds et al., 2018). The addition of an NP decline marker to the existing protocols could aid in the identification and recruitment of high-risk participants for clinical trials focusing on preclinical or MCI populations.

Many prospective studies of aging follow participants with annual or semi-annual neuropsychological exams, but these data are not always used to determine dementia risk. The NP decline approach presented earlier provides simple equations for standardizing the discrepancy between expected performance and actual performance at follow-up



No. at risk								
Normal / NP-	0	39	35	30	26	13	6	1
Normal / NP+	0	40	26	23	11	10	4	0
Impaired MCI- / NP-	0	12	8	5	3	1	1	0
Impaired MCI- / NP+	0	12	6	6	2	0	1	0
Impaired MCI+ / NP-	0	74	37	28	6	4	3	0
Impaired MCI+ / NP+	1	164	73	30	15	7	1	0

FIGURE 2 | Progression to dementia stratified by cognitive status and NP decline status in the National Alzheimer's Coordinating Center Database. Cumulative progression to dementia from Cox regression analysis is displayed and stratified by baseline NACC diagnosis, including Normal Cognition (Normal), Impaired without MCI (Impaired MCI−), Impaired with MCI (Impaired MCI+), and NP decline status at 12-month follow-up based on optimal cutoff values, including NP decline absent (NP−, above 28th percentile) and NP decline present (NP+, at or below 28th percentile). The table below displays the number of participants who progressed to dementia at each follow-up interval.

(Crawford and Garthwaite, 2006; Slick, 2006; Nation et al., 2019). The NP decline metric may be valuable in the context of these longitudinal aging studies since 12-month NP decline can be easily calculated to determine whether participants are showing worse than expected follow-up performance. Critically, participants showing NP decline beyond optimal cutoff values were at an increased risk for future dementia even if they were still performing within the normative range at 12-month follow-up. Clinicians often follow at-risk individuals on an annual or semi-annual basis, yielding serial neuropsychological data that can be easily evaluated using the provided equations and cutoff values for NP decline quantification.

Data from 12-month NP decline may help inform clinician judgments since decline beyond optimal cutoff values has now been linked to an approximately 2-fold increase in risk for dementia in two large longitudinal cohorts (Nation et al., 2019). Thus, there may be immediate value in terms of both research and clinical applications of the NP decline metric, allowing clinicians to gather further prognostic information beyond that obtained by the diagnosis of normal cognition or MCI. It is also important to note that even short-term practice effects (e.g., exams separated by 1 week) have also shown to be indicative of later cognitive decline (e.g., Duff et al., 2011). Practice effects across 1 week are related to diagnosis (Duff et al., 2008), prognosis (Duff et al., 2007, 2011), and treatment response (Duff et al., 2010), showing how the examination of these is another critical future direction of this work.

The potential application of NP decline analysis goes beyond any specific dementia etiology, but it should also be noted that recent research recommendations for the diagnosis of Alzheimer's disease have emphasized the evaluation of serial cognitive test data to determine early or subtle cognitive decline (Jack et al., 2018). Although prior study has focused primarily on single exam methods for identifying older adults with subtle cognitive decline (Donohue et al., 2014; Edmonds et al., 2015b; Toledo et al., 2015), serial exams may be required in order to detect the earliest cognitive changes represented by a decline within normal range performance. The method employed in this study allows for quantification and standardization of longitudinal decline within normal range performance, which may better detect subtle cognitive changes related to an incipient neuropathological process. Numerous studies have emphasized the role of biomarkers in the stratification of dementia risk in cognitively unimpaired older adults (Jack et al., 2018), but other studies have shown that many older adults with biomarker abnormalities will never develop dementia (Ritchie et al., 2017). Combining sensitive preclinical neuropsychological instruments with preclinical biomarkers may aid in prognostic evaluation and treatment decision-making beyond information obtained through biomarker analysis alone (Nation et al., 2019).

Strengths of this study include the longitudinal analysis and large sample size. Limitations include the variable clinical follow-up and heterogeneity of NACC sampling methods that includes a mixture of studies from numerous sites with both clinical- and community-based studies. Furthermore, the NACC database has limited ethnic diversity, with NACC participants being largely Caucasian. However, of note, the NACC database does enroll participants with diverse medical history, including dementia of various etiologies, and this heterogeneity of NACC data benefits the generalizability of the study findings, particularly since the results coincided well with the recently published data from the more curated ADNI study sample (Nation et al., 2019). The use of neuropsychological test data to predict future dementia risk has also been criticized for circularity. Although neuropsychological test data can often be used to aid in the diagnosis of dementia in conjunction with other data, including measures of functional decline, informant reports, behavioral observations, and clinician judgments, this study evaluated the predictive value of neuropsychological markers in older adults with normal to mildly impaired cognitive function. Thus, neuropsychological markers may be useful prognostic instruments capable of stratifying future dementia risk even in patients with normative cognition, or only mild cognitive changes, with no functional decline or very minimal functional change. In this context, neuropsychological markers are not diagnostic of dementia, but rather they are prognostic indicators that may be of value in the detection of an incipient decline in neurocognitive function, potentially presaging the future development of major cognitive and functional impairments that characterize dementia. The use of cognitive data to predict dementia risk based on MCI diagnosis is a well-established practice (Petersen, 2011) that is no more circular than the use of neuropsychological markers to predict future dementia from an even earlier stage, as in this study. Just as MCI is a risk factor for dementia, NP decline is a risk factor for dementia. These risk factors are not circular. One of the most valuable aspects of NP decline is that it may be used in conjunction with MCI diagnosis, or even in cognitively unimpaired older adults, further stratifying and refining dementia risk assessment.

Additional research and development of methods for longitudinal analysis of serial neuropsychological exam data will improve our ability to determine patient cognitive trajectories, which will have major implications for neuropsychological research, clinical trials, and clinical practice in a variety of patient populations.

DATA AVAILABILITY STATEMENT

The original contributions presented in this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

DN drafted the manuscript and acquired financial support for this manuscript. JH and DN developed and designed the study, conducted analyses and interpretation of the data, revised the manuscript, and approved the submitted version.

FUNDING

This NACC database was funded by the NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer), P30 AG066468 (PI Oscar Lopez), P30 AG062421 (PI Bradley Hyman), P30 AG066509 (PI Thomas Grabowski), P30 AG066514 (PI Mary Sano), P30 AG066530 (PI Helena Chui), P30 AG066507 (PI Marilyn Albert), P30 AG066444 (PI John Morris), P30 AG066518 (PI Jeffrey Kaye), P30 AG066512 (PI Thomas Wisniewski), P30 AG066462 (PI Scott Small), P30 AG072979 (PI David Wolk), P30 AG072972 (PI Charles DeCarli), P30 AG072976 (PI Andrew Saykin), P30 AG072975 (PI David Bennett), P30 AG072978 (PI Neil Kowall), P30 AG072977 (PI Robert Vassar), P30 AG066519 (PI Frank LaFerla), P30 AG062677 (PI Ronald Petersen), P30 AG079280 (PI Eric Reiman), P30 AG062422 (PI Gil Rabinovici), P30 AG066511 (PI Allan Levey), P30 AG072946 (PI Linda Van Eldik), P30 AG062715 (PI Sanjay Asthana, FRCP), P30 AG072973 (PI Russell Swerdlow), P30 AG066506 (PI Todd Golde), P30 AG066508 (PI Stephen Strittmatter), P30 AG066515 (PI Victor Henderson), P30 AG072947 (PI Suzanne Craft),

P30 AG072931 (PI Henry Paulson), P30 AG066546 (PI Sudha Seshadri), P20 AG068024 (PI Erik Roberson), P20 AG068053 (PI Justin Miller), P20 AG068077 (PI Gary Rosenberg), P20 AG068082 (PI Angela Jefferson), P30 AG072958 (PI Heather Whitson), and P30 AG072959 (PI James Leverenz).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.838459/full#supplementary-material

REFERENCES

Amieva, H., Jacqmin-Gadda, H., Orgogozo, J. M., Le Carret, N., Helmer, C., Letenneur, L., et al. (2005). The 9 year cognitive decline before dementia of the Alzheimer type: a prospective population-based study. *Brain* 128(Pt 5), 1093–1101. doi: 10.1093/brain/awh451

Binder, L. M., Iverson, G. L., and Brooks, B. L. (2009). To err is human: "abnormal" neuropsychological scores and variability are common in healthy adults. *Arch. Clin. Neuropsychol.* 24, 31–46. doi: 10.1093/arclin/acn001

Bondi, M. W., Jak, A. J., Delano-Wood, L., Jacobson, M. W., Delis, D. C., and Salmon, D. P. (2008). Neuropsychological contributions to the early identification of Alzheimer's disease. *Neuropsychol. Rev.* 18, 73–90. doi: 10.1007/s11065-008-9054-1

Crawford, J. R., and Garthwaite, P. H. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: a significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology* 20, 259–271. doi: 10.1037/0894-4105.20.3.259

Crawford, J. R., and Howell, D. C. (1998). Regression equations in clinical neuropsychology: an evaluation of statistical methods for comparing predicted and obtained scores. *J. Clin. Exp. Neuropsychol.* 20, 755–762. doi: 10.1076/jcen.20.5.755.1132

Crous-Bou, M., Minguillon, C., Gramunt, N., and Molinuevo, J. L. (2017). Alzheimer's disease prevention: from risk factors to early intervention. *Alzheimers Res. Ther.* 9:71. doi: 10.1186/s13195-017-0297-z

Donohue, M. C., Sperling, R. A., Salmon, D. P., Rentz, D. M., Raman, R., Thomas, R. G., et al. (2014). The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurol.* 71, 961–970. doi: 10.1001/jamaneurol.2014.803

Duff, K., Beglinger, L. J., Moser, D. J., Paulsen, J. S., Schultz, S. K., and Arndt, S. (2010). Predicting cognitive change in older adults: the relative contribution of practice effects. *Arch. Clin. Neuropsychol.* 25, 81–88.

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., Haase, R. F., et al. (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch. Clin. Neuropsychol.* 22, 15–24. doi: 10.1016/j.acn.2006.08.013

Duff, K., Beglinger, L. J., Van Der Heiden, S., Moser, D. J., Arndt, S., Schultz, S. K., et al. (2008). Short-term practice effects in amnestic mild cognitive impairment: implications for diagnosis and treatment. *Int. Psychogeriatr.* 20, 986–999. doi: 10.1017/S1041610208007254

Duff, K., Hammers, D. B., Dalley, B. C. A., Suhrie, K. R., Atkinson, T. J., Rasmussen, K. M., et al. (2017). Short-term practice effects and amyloid deposition: providing information above and beyond baseline cognition. *J. Prev. Alzheimers Dis.* 4, 87–92. doi: 10.14283/jpad.2017.9

Duff, K., Lyketsos, C. G., Beglinger, L. J., Chelune, G., Moser, D. J., Arndt, S., et al. (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 19, 932–939.

Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., et al. (2015a). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimers Dement.* 11, 415–424. doi: 10.1016/j.jalz.2014.03.005

Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., and Bondi, M. W. (2015b). Subtle cognitive decline and biomarker staging in preclinical Alzheimer's disease. *J. Alzheimers Dis.* 47, 231–242. doi: 10.3233/JAD-150128

Edmonds, E. C., Delano-Wood, L., Galasko, D. R., Salmon, D. P., Bondi, M. W., and Alzheimer's Disease Neuroimaging Initiative (2014). Subjective cognitive complaints contribute to misdiagnosis of mild cognitive impairment. *J. Int. Neuropsychol. Soc.* 20, 836–847. doi: 10.1017/S135561771400068X

Edmonds, E. C., Weigand, A. J., Thomas, K. R., Eppig, J., Delano-Wood, L., Galasko, D. R., et al. (2018). Increasing inaccuracy of self-reported subjective cognitive complaints over 24 months in empirically derived subtypes of mild cognitive impairment. *J. Int. Neuropsychol. Soc.* 24, 842–853. doi: 10.1017/S1355617718000486

Han, S. D., Nguyen, C. P., Stricker, N. H., and Nation, D. A. (2017). Detectable neuropsychological differences in early preclinical Alzheimer's disease: a meta-analysis. *Neuropsychol. Rev.* 27, 305–325. doi: 10.1007/s11065-017-9345-5

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Ho, J. K., and Nation, D. A., (2018). Neuropsychological profiles and trajectories in preclinical Alzheimer's disease. *J. Int. Neuropsychol. Soc.* 24, 693–702. doi: 10.1017/S135561771800022X

Jack, C. R. Jr., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 14, 535–562. doi: 10.1016/j.jalz.2018.02.018

Koscik, R. L., Jonaitis, E. M., Clark, L. R., Mueller, K. D., Allison, S. L., Gleason, C. E., et al. (2019). Longitudinal standards for mid-life cognitive performance: identifying abnormal within-person changes in the wisconsin registry for Alzheimer's prevention. *J. Int. Neuropsychol. Soc.* 25, 1–14. doi: 10.1017/S1355617718000929

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., DeCarli, C., Ferris, S., et al. (2006). The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer disease centers. *Alzheimer Dis. Assoc. Disord.* 20, 210–216.

Nation, D. A., Ho, J. K., Dutt, S., Han, S. D., Lai, M. H. C., and Alzheimer's Disease Neuroimaging Initiative (2019). Neuropsychological decline improves prediction of dementia beyond Alzheimer's disease biomarker and mild cognitive impairment diagnoses. *J. Alzheimers Dis.* 69, 1171–1182. doi: 10.3233/JAD-180525

Papp, K. V., Buckley, R., Mormino, E., Maruff, P., Villemagne, V. L., Masters, C. L., et al. (2020). Clinical meaningfulness of subtle cognitive decline on longitudinal testing in preclinical AD. *Alzheimers Dement.* 16, 552–560. doi: 10.1016/j.jalz.2019.09.074

Petersen, R. C. (2011). Clinical practice. Mild cognitive impairment. *N. Engl. J. Med.* 364, 2227–2234. doi: 10.1056/NEJMcp0910237

Ritchie, C., Smailagic, N., Noel-Storr, A. H., Ukoumunne, O., Ladds, E. C., and Martin, S. (2017). CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochr. Database Syst. Rev.* 3:CD010803. doi: 10.1002/14651858.CD010803.pub2

Slick, D. J. (2006). "Psychometrics in neuropsychological assessment," in *A Compendium of Neuropsychological Tests*, Third Edn, eds E. M. S. Sherman and O. Spreen (New York, NY: Oxford University Press), 3–31.

Thomas, K. R., Bangen, K. J., Weigand, A. J., Edmonds, E. C., Wong, C. G., Cooper, S., et al. (2020). Objective subtle cognitive difficulties predict future amyloid accumulation and neurodegeneration. *Neurology* 94, e397–e406. doi: 10.1212/WNL.0000000000008838

Thomas, K. R., Edmonds, E. C., Eppig, J. S., Wong, C. G., Weigand, A. J., Bangen, K. J., et al. (2019). MCI-to-normal reversion using neuropsychological criteria in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement.* 15, 1322–1332. doi: 10.1016/j.jalz.2019.06.4948

Toledo, J. B., Bjerke, M., Chen, K., Rozycki, M., Jack, C. R. Jr., Weiner, M. W., et al. (2015). Memory, executive, and multidomain subtle cognitive impairment:

clinical and biomarker findings. *Neurology* 85, 144–153. doi: 10.1212/WNL.0000000000001738

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 28 July 2022
DOI 10.3389/fnagi.2022.911559

OPEN ACCESS

EDITED BY
Lars Nyberg,
Umeå University, Sweden

REVIEWED BY
Jeremy Andrew Elman,
University of California, San Diego, United States
Michael Malek-Ahmadi,
Banner Alzheimer's Institute, United States

*CORRESPONDENCE
Andrew R. Bender
arbender@msu.edu

SPECIALTY SECTION
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 02 April 2022
ACCEPTED 27 June 2022
PUBLISHED 28 July 2022

CITATION
Bender AR, Ganguli A, Meiring M, Hampstead BM and Driver CC (2022) Dynamic modeling of practice effects across the healthy aging-Alzheimer's disease continuum. *Front. Aging Neurosci.* 14:911559. doi: 10.3389/fnagi.2022.911559

Dynamic modeling of practice effects across the healthy aging-Alzheimer's disease continuum

Andrew R. Bender[1,2,3]*, Arkaprabha Ganguli[4], Melinda Meiring[2], Benjamin M. Hampstead[3,5,6] and Charles C. Driver[7,8]

[1]Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, United States, [2]Graduate Program in Neuroscience, College of Natural Science, Michigan State University, East Lansing, MI, United States, [3]Michigan Alzheimer's Disease Research Center, Ann Arbor, MI, United States, [4]Department of Statistics and Probability, College of Natural Science, Michigan State University, East Lansing, MI, United States, [5]Mental Health Service, VA Ann Arbor Healthcare System, Ann Arbor, MI, United States, [6]Research Program on Cognition and Neuromodulation Based Intervention, Department of Psychiatry, University of Michigan, Ann Arbor, MI, United States, [7]Institute of Education, University of Zurich, Zurich, Switzerland, [8]Institute for Educational Evaluation, Associated Institute at the University of Zurich, Zurich, Switzerland

Standardized tests of learning and memory are sensitive to changes associated with both aging and superimposed neurodegenerative diseases. Unfortunately, repeated behavioral test administration can be confounded by practice effects (PE), which may obscure declines in level of abilities and contribute to misdiagnoses. Growing evidence, however, suggests PE over successive longitudinal measurements may differentially predict cognitive status and risk for progressive decline associated with aging, mild cognitive impairment (MCI), and dementia. Thus, when viewed as a reflection of neurocognitive plasticity, PE may reveal residual abilities that can add to our understanding of age- and disease-related changes in learning and memory. The present study sought to evaluate differences in PE and verbal recall in a clinically characterized aging cohort assessed on multiple occasions over 3 years. Participants included 256 older adults recently diagnosed as cognitively unimpaired (CU; n = 126), or with MCI of amnestic (n = 65) or non-amnestic MCI (n = 2085), and multi-domain amnestic dementia of the Alzheimer's type (DAT; n = 45). We applied a continuous time structural equation modeling (ctsem) approach to verbal recall performance on the Hopkins Verbal Learning Test in order to distinguish PE from individual occasion performance, coupled random changes, age trends, and differing measurement quality. Diagnoses of MCI and dementia were associated with lower recall performance on all trials, reduced PE gain per occasion, and differences in non-linear dynamic parameters. Practice self-feedback is a dynamic measure of the decay or acceleration in PE process changes over longitudinal occasions. As with PE and mean recall, estimated practice self-feedback followed a gradient from positive in CU participants to null in participants with diagnosed MCI and negative for those with dementia

diagnoses. Evaluation of sensitivity models showed this pattern of variation in PE was largely unmodified by differences in age, sex, or educational attainment. These results show dynamic modeling of PE from longitudinal performance on standardized learning and memory tests can capture multiple aspects of behavioral changes in MCI and dementia. The present study provides a new perspective for modeling longitudinal change in verbal learning in clinical and cognitive aging research.

KEYWORDS

practice effects, aging, learning, mild cognitive impairment, verbal memory, dementia, dynamic modeling, Alzheimer's disease (AD)

Introduction

Standardized neuropsychological tests are sensitive to cognitive declines associated with older age and incident mild cognitive impairment (MCI) and dementia. Clinical characterization of cognitive impairments and the tracking of progressive declines requires repeated testing, but performance on repeated standardized tests is contaminated by practice effects (PE; Duff et al., 2001; Hawkins et al., 2004; Salthouse, 2010; Hoffman et al., 2011). This contamination arises due to the incidental retention of information from prior exposure to test format and content, which can enhance performance at subsequent reinstatement (Wilson et al., 2000; Heilbronner et al., 2010; Machulda et al., 2013). The potential of PE to mask true cognitive declines in healthy and pathological aging has motivated numerous attempts to remove PE from estimates of level or change in performance (Rabbitt et al., 2001; Salthouse and Tucker-Drob, 2008; Salthouse, 2010; Calamia et al., 2012). However, as a measure of the capacity to benefit from repetition, PE may represent an independent behavioral dimension sensitive to declines in older age and neurodegenerative disease (Yang, 2011; Duff et al., 2012). Thus, rather than treat PE as noise, approaches to integrate modeling of PE and cognition may provide novel clinical value in characterizing cognitive impairment and dementia.

Notably, simulation study findings show PE are not easily distinguished from true changes associated with aging or cohort effects (Hoffman et al., 2011). Therefore, quantifying PE as the change dimension of interest may better serve short-term characterization of functional declines in MCI and dementia. This proposition is in accord with suggestions that variation in PE reflects individual differences in neurocognitive plasticity (Baltes and Raykov, 1996; Yang and Krampe, 2009; Yang, 2011). Others have reported PE as a marker of clinical declines in older adults with mild cognitive impairment (MCI) or dementia of the Alzheimer's type (DAT; Duff et al., 2007, 2012; Fernandez-Ballesteros et al., 2012; Sanchez-Benavides et al., 2016). Lower PE is also associated with performance decrements in cognitively intact adults with preclinical Alzheimer's pathology (Goldberg et al., 2015; Hassenstab et al., 2015). These findings highlight the intrinsic dependencies between the contributions of prior experience to cognitive performance and vulnerability to decline.

Verbal learning tasks provide established clinical markers of neuropsychological deficits associated with diagnoses of MCI and DAT (Duff et al., 2001; Hawkins et al., 2004; Blasi et al., 2009; Lonie et al., 2010; Summers and Saunders, 2012). Standardized tests of verbal learning and memory typically involve serial auditory presentation of lists of verbal stimuli, immediately followed by instructions to freely recall all words remembered. Most standardized tasks then repeat this procedure for multiple trials with the same stimuli, followed by a delay and an additional free recall trial. Due to their repetitive nature, verbal learning tasks are particularly vulnerable to PE when content is repeated across longitudinal administrations (Duff et al., 2001; Heilbronner et al., 2010; Machulda et al., 2013; Campos-Magdaleno et al., 2017). Arguably, repeated free recall performance on multiple trials distributed over longitudinal occasions embodies the definition of a dynamic process – i.e., one that constantly changes and progresses (Zimprich et al., 2008). Moreover, serial recall represents retrieval-based learning, in which retrieval of a representation updates the representation itself (Karpicke et al., 2014). Furthermore, each repeated trial involves not just encoding and retrieval, but updating and retrieval monitoring, as well as potential metacognitive processes (Hertzog and Dunlosky, 2004; Bender and Raz, 2012). Thus, successful recall performance involves multiple interactive executive processes, which may also show decrements in the presence of phenotypic cognitive impairment. Yet, the extent that task summary scores reflect these dynamics is unclear.

To date, there is neither consistent operationalization nor definition of PE in the contexts of clinical and basic cognitive aging research. Studies report PE estimated both in variable time scales ranging from minutes to years and from a host of different behavioral tasks, conditions, and

stimuli. In addition, PE is largely quantified in extant studies of normative and pathological aging using difference scores or *via* linear modeling frameworks (Raykov et al., 2002; Salthouse et al., 2004; Duff et al., 2007; Bender et al., 2013, 2020; Goldberg et al., 2015; Hassenstab et al., 2015). However, linear modeling approaches may fail to capture the interactive, dynamic processes involved in longitudinal verbal learning task performance. Modern dynamic modeling methods that can quantify non-linear processes may provide novel markers of PE or cognitive decline. In the context of longitudinal changes in verbal recall, dynamic modeling can account for the current level of performance at each trial and occasion to help predict subsequent performance. Thus, within-occasion and longitudinal performance are modeled as interdependent processes that play out over time. Modeling performance on each verbal recall trial as an individual interactive process, manifest over multiple occasions, permits estimating PE as a change process independent of overall mean performance and trial-by-trial random effects.

The continuous time structural equation modeling (*ctsem*; Driver et al., 2017) framework applies a differential equations-based time series analysis for modeling ongoing dynamic processes, coupled with a measurement layer to delineate measurement noise from true change. While it resembles latent growth and latent change score models, *ctsem* permits treating time-in-study as a continuous variable, in addition to other key enhancements. Relevant to longitudinal verbal learning performance, the framework permits specifying a non-linear measurement model to account for factors such as differential measurement error across groups or levels of performance. It also allows modeling random effects to capture individual differences in all system parameters, as well as covariates that can predict such individual differences. Furthermore, these non-linear processes may also be sensitive to phenotypic cognitive impairment, possibly independent of level of performance or PE. Thus, dynamic modeling of longitudinal verbal learning data to decompose PE from trial-level performance may offer additional value for clinical aging populations previously reported to show a loss of PE.

Extant findings show diagnoses of MCI and multi-domain amnestic dementia (DAT) are associated with reduced or non-existent PE on standardized verbal learning tasks (Duff et al., 2007, 2019; Calamia et al., 2012; Goldberg et al., 2015; Gavett et al., 2016). However, it is unclear whether differences in PE associated with MCI or dementia are also influenced by other factors known to influence verbal memory. For example, whereas older age is associated with performance decrements on episodic memory tasks, female sex is associated with better verbal episodic memory (Herlitz et al., 1997; Herlitz and Rehnman, 2008; Bender et al., 2010). Furthermore, greater educational attainment also confers a higher initial level of premorbid performance on memory tasks (Lovden et al., 2020). Still, it is unclear if such individual differences may modify the larger effects of MCI or DAT diagnosis on verbal recall or PE, particularly over less expansive periods of assessment.

The University of Michigan Memory and Aging Project (UM-MAP) includes older participants clinically characterized as cognitively unimpaired (CU) or diagnosed with MCI or DAT. The available data includes one to four occasions of annual neuropsychological assessment, including the Hopkins Verbal Learning Test (HVLT), which was administered using the same stimulus lists on each occasion of measurement. This provided an opportunity to apply *ctsem* for modeling longitudinal verbal recall performance and PE as dynamic processes in a clinical aging sample. To our knowledge, this is the first attempt to apply dynamic modeling to quantify longitudinal changes in verbal learning, particularly in a clinical aging context. Critically, dynamically modeled estimates of PE in the present study served as the primary measure of longitudinal change in performance, rather than estimating change in ability and PE separately. We hypothesized that both older age and diagnosed MCI and DAT would be associated with poorer recall and lower PE. We also expected the effects of clinical diagnosis would be modified by individual differences in chronological age, sex, and educational attainment. Specifically, we hypothesized that higher education and female sex would be associated with better verbal recall; however, we had no clear expectations regarding how these would influence effects of MCI or DAT diagnosis on recall or PE.

Materials and methods

Participants

The study sample was drawn from research participants in the University of Michigan Memory and Aging Project (UM-MAP), which is the primary clinical cohort at the Michigan Alzheimer's Disease Research Center (MADRC). The sample included 256 participants (67% women) from 51 to 89 years of age at the first assessment. At each measurement occasion all participants underwent neuropsychological evaluation and a consensus diagnosis was made during a consensus conference by neurologists, neuropsychologists, nurses, social workers or other specialists as appropriate using the National Alzheimer's Coordinating Center (NACC) criteria. The sample was divided into three subgroups based on the last recorded diagnosis for each participant (**Table 1**): cognitively unimpaired (CU; n = 126; 71% women), amnestic or non-amnestic MCI (MCI; n = 85; 67% women) and multi-domain amnestic dementia (DAT; n = 45; 60% women) consistent with Alzheimer's disease and mixed dementia. Over the course of the study, six participants progressed from diagnoses of aMCI to DAT of the Alzheimer's type, and an additional six participants changed from CU to MCI diagnoses. In contrast, one participant initially diagnosed with DAT was subsequently

characterized as CU, and 15 participants characterized with MCI at baseline reverted to CU at their final study assessment 2–3 years later.

Longitudinal organization

Following baseline assessment participants returned annually for repeated testing. The present study data included assessments on one to four separate measurement occasions (Table 2), separated by mean intervals of 1.09 years. Mean intervals between each occasion of measurement, separately by subgroup are reported in Supplementary Table 1.

Cognitive testing

All participants were administered the Hopkins Verbal Learning Test (HVLT; Brandt, 1991) on each occasion of measurement and testing followed the published procedures. The HVLT auditorily presents 12-item lists of semantically linked verbal stimuli, presented at a rate of 2 s per item. Following presentation of all items, the participant freely recalls as many as possible. The score per trial is the total number of correctly recalled words. This is repeated for two additional free recall trials, using the same verbal stimuli. A 20-min delay follows the third recall trial, after which participants are asked to freely recall as many words as possible without re-presenting the stimuli. Notably, although the HVLT also includes additional delayed recall and recognition measures, the present study focused on the first four trials, i.e., the three immediate and first delayed recall trials. Critically, the present study repeated the same lists of verbal stimuli across all occasions of measurement.

TABLE 1 Participant characteristics by clinical diagnosis.

	CU	MCI	DAT
	Mean (*sd*)	Mean (*sd*)	Mean (*sd*)
Age	70.06 (6.43)	72.66 (8.04)	72.22 (9.31)
Education	15.90 (2.67)	15.60 (2.47)	15.51 (2.61)
Systolic BP	134.43 (22.97)	139.46 (22.10)	133.81 (15.67)
Diastolic BP	77.88 (11.35)	81.10 (12.03)	74.91 (9.76)
CDR	0.33 (0.41)	0.97 (0.75)	3.67 (2.08)
MoCA	26.85 (1.93)	23.03 (3.19)	15.45 (5.64)
GDS	1.13 (1.39)	1.43 (1.70)	1.60 (1.33)

Values are mean with standard deviation in parentheses. CU, cognitively unimpaired; MCI, diagnosis of amnestic or non-amnestic MCI; DAT, diagnosis of multi-domain amnestic dementia. Age and educational attainment are in years. Systolic and diastolic BP are blood pressure measured in mmHg. CDR, clinical dementia rating; MoCA, montreal cognitive assessment. GDS, geriatric depression scale.

TABLE 2 Participant counts for total number of measurement occasions by clinical diagnosis.

Clinical diagnosis	Total number of occasions				
	1	2	3	4	Total
CU	20	45	40	21	126
MCI	34	23	20	8	85
DAT	20	19	6	0	45
Total	74	87	66	29	256

CU, cognitively unimpaired; MCI, diagnosis of amnestic or non-amnestic MCI; DAT, diagnosis of multi-domain amnestic dementia. Values represent counts of participants by their total number of measurement occasions. For example, in the top row for CU participants, 20 had HVLT data for only one occasion, 45 participants completed two occasions, 40 had three occasions of data, and 21 CU participants had complete data for all four occasions.

Data analysis

To analyze performance, change as a function of PE, and individual differences therein, we developed hierarchical Bayesian continuous time dynamic models (Driver et al., 2017) implemented in the *ctsem* software (Driver et al., 2017;Driver and Voelkle, 2021). A more detailed description of the model and corresponding mathematical apparatus follows below in the Supplementary Material.

Modeling practice effects and performance in *ctsem*

To account for varying observation timing and to allow for continuously interacting processes, *ctsem* estimates an underlying continuous-time model, which is translated into discrete time expectations and covariance matrices using matrix exponentiation (Voelkle et al., 2012; Voelkle and Oud, 2013). To account for the multiple timescales at play (i.e., within and between occasion), each of the immediate (i.e., Trials 1, 2, and 3) and delayed recall trials (Trial 4) were modeled as independent latent processes over four occasions of measurement, with correlated random disturbances. This means that although we may not have been able to predict every fluctuation in performance, when an unpredicted fluctuation occurs this contributes to predictions for the other trials within and (potentially) across occasion. We estimated the standard deviation and within-occasion correlations of the diffusion process, separately for each Trial (e.g., Diffusion T1). These parameters capture the extent of unpredictable random changes across measurement occasions, which are nevertheless useful for predicting performance on other trials within-occasion, or across-occasion – thus more likely representing some genuine aspect of performance. In contrast, the standard deviation of the measurement error (i.e., *measurement error*) captures unpredictable changes in observed performance that do not provide value for prediction on other trials. The model also contained a parameter reflecting *Trial self-feedback* (*sf_Trial*); this parameter describes the persistence of the random changes

between measurement occasions for each trial. Put differently, sf_Trial represents the extent to which unpredicted shifts up or down (independent of measurement error) on performance for a specific trial, can be used to predict performance for the same trial number on the next occasion, i.e., across-occasion persistence. On top of this base structure allowing for correlated random processes, PE was modeled as a latent process that changed at the end of each occasion. As for the trial specific processes, we also specified a Practice self-feedback (sf_Practice) parameter to provide an estimate of total feedback on PE; this parameter serves as a measure of the decay or acceleration in the change to Practice effect process over the observed range of occasions. Like sf_Trial, the sf_Practice parameter reflects the extent that the current level of practice (i.e., at the end of each occasion) contributes to the Practice effect at the next occasion. Thus, a positive sf_Practice value would reflect an increase in gains due to practice on later occasions, whereas a negative sf_Practice value reflects a decay or deceleration of learning processes that reduce total PE and implies reducing gains due to further practice. Each model output includes estimates of population means for all modeled parameters, correlations between trial manifest means and the PE parameter, and estimates of time independent predictor effects and interactions.

Time independent predictors and covariates

Parameters of the system and measurement models also varied on an individual level as a function of clinical diagnosis, as well as random effects. The effects of other sources of individual differences were examined in separate sensitivity models to evaluate effect modification by individual linear covariates, including baseline age, sex, and educational attainment. This accounts for a broad range of phenomena, such as heterogeneity of measurement error variance with age and performance. Therefore, we first evaluated most recent clinical diagnosis of MCI or DAT as time independent predictors, in relation to CU participants. This was followed by independent subsidiary sensitivity models that evaluated covariate effects age, sex, and educational attainment (scaled and centered at the respective sample means) on model parameters and their interactions with diagnostic group. Last, independently for the three diagnostic groups we evaluated each of the time independent predictors age, sex, and education in separate models.

Bayesian estimation

Due to the large number of parameters and random effects, we opted for Bayesian maximum *a posteriori* estimation. Priors on the parameter means were relatively broad and non-influential, while tighter priors (i.e., pushing estimates toward zero) were used for modeling individual differences to mitigate over-fitting. Despite yielding more conservative estimates, this permits a more pragmatic approach for estimating and interpreting models with many parameters and modest sample sizes. For details on priors, and the expanded stochastic differential equation and related measurement model see the **Supplementary Material**.

Results

A guide to interpreting model results and figures

The time independent predictor effects and interactions are best represented by the accompanying expectation plots (**Figures 1–4**). As these are likely to be unfamiliar to most readers without prior dynamic modeling experience, their interpretation benefits from some explanation. **Figure 1** provides an example of the expected effects of educational attainment on performance, uncomplicated by additional interactions. The plot depicts model expectations of recall performance, measured over four occasions, with each trial type (i.e., 1–4) depicted separately in the four panels. The *y*-axes represent the number of correctly recalled words on a trial, and the *x*-axis represents time; the dashed vertical lines depict the individual measurement occasions. The black plotted line depicts the expectation of change in the total sample, in the absence of any covariates. The level of the line on the *y*-axis represents the number of words recalled for a trial in the total sample and this expected value is incrementally increased before the next measurement occasion as a function of the estimated PE parameter. Starting at baseline (T0) the line is flat until just before the second occasion (T1) where PE is first relevant. The magnitude of the increase reflects PE at that occasion. The slope of the line between T1 to T2 and from T2 to T3 also reflects the amount of positive feedback or decay in PE as estimated by positive or negative sf_Practice – the feedback component on PE that allows for increasing or reducing gains of further practice. The dashed and solid red lines show model expectations when the covariate in question is ± 1 and all other covariates are zero. For dichotomous covariates like sex, this reflects group differences. Here, higher education (dashed red line) predicts higher level of performance, with stronger effects on Trials 2–4 and occasions T0–T2, but the difference is reduced at T3. The lower education group (solid red line) has a higher PE gain at the end of each occasion, despite the lower initial level. In addition, this is accompanied by greater decay (i.e., less positive practice self-feedback) on the PE process. Of note, for interactions (e.g., **Figures 2–4**), the dashed or solid lines represent only the interaction effect, not the interaction plus main effects. For example, in the case of Sex $\times$ DAT, the $+1$ line shows only the additive effect of female sex and positive DAT diagnosis, assuming the individual Sex and DAT diagnosis covariates are 0.

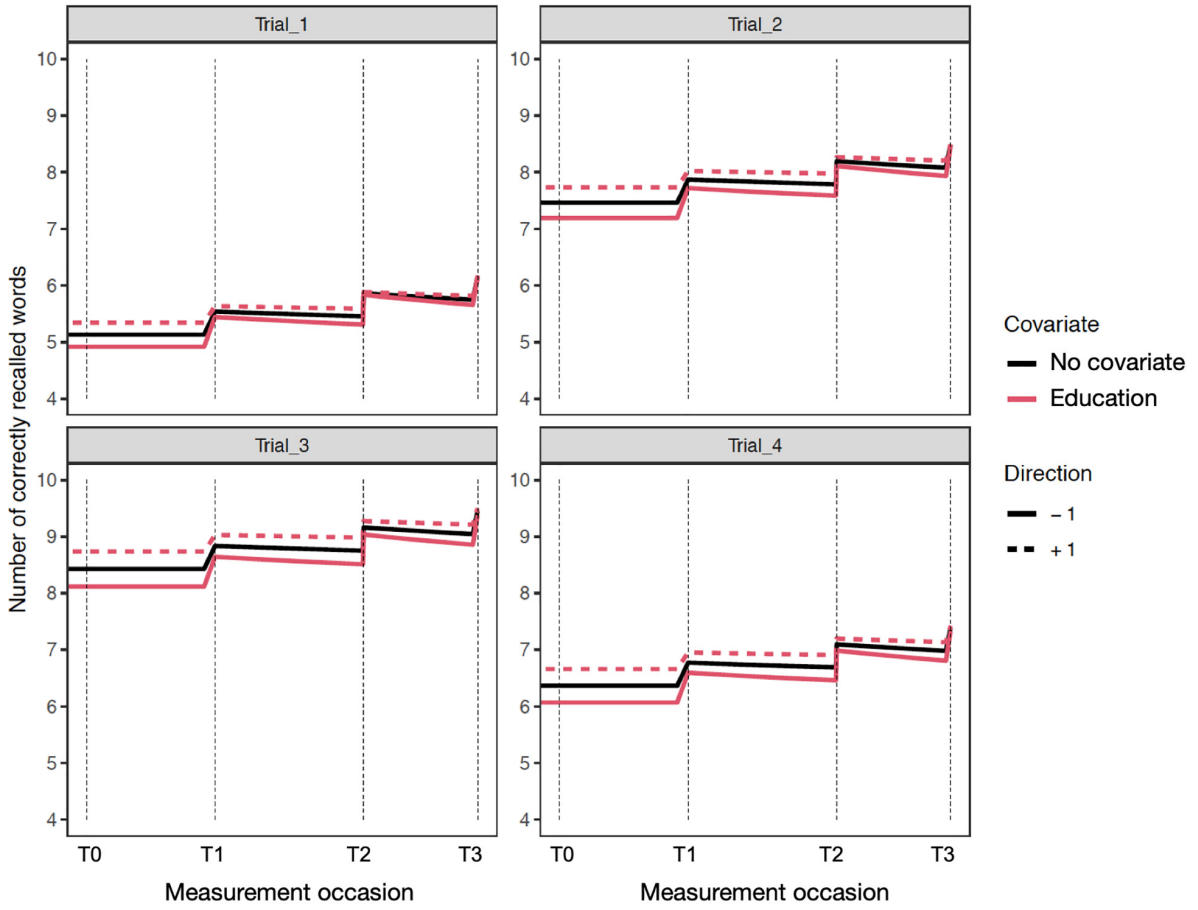


FIGURE 1
Expectation plots for change in recall performance (black line) across four measurement occasions (e.g., T0: baseline, T1: 1 year). The solid and dashed red lines depict effect modification by education; higher (dashed) and lower (solid) levels of education predict different levels and patterns of change. The complete guide to interpreting expectation plots can be found in Section "A Guide to Interpreting Model Results and Figures."

Diagnostic groups model

The model with diagnostic groups as the only time independent predictor provided overall characterization of the sample (Table 3). The estimated population mean of PE was significantly positive, indicating overall improvement in performance across longitudinal occasions of 0.4 words, for the entire sample. The mean trial self-feedback (sf_Trial) parameter was very negative, implying that random changes at the trial level did not persist across occasions. In addition, the mean for the sf_Practice parameter was not significant, suggesting that gains in PE neither increase nor decrease substantially, given further repetitions. The means for the other parameters, including diffusion for each trial and total measurement error were all positive. Furthermore, the means for all four Trials were positively correlated within-subject, but there were no significant correlations with PE gain per occasion in the total sample. Evaluation of diagnostic groups as time independent predictors showed MCI and dementia diagnosis predicted lower mean performance on all four trials (Table 4), as well as a non-significant trend for DAT diagnosis predicting lower PE. Both MCI and DAT diagnoses predicted significantly higher Diffusion effects for Trial 3 only, implying that diagnoses of MCI and dementia were associated with greater random changes in Trial 3 that were nevertheless predictive of other trials, thus likely representing genuine change and not measurement error.

Sensitivity models

Next, we evaluated independent sensitivity models to examine the modifying effects of individual differences in age, sex, and educational attainment on main effects and interactions with diagnostic group (Table 5).

TABLE 3 Estimated population mean values and correlations for diagnostic group and sensitivity models.

	Diagnostic groups		Age		Sex		Education	
	Mean (*SD*)	95% CI	Mean (*SD*)	95% CI	Mean (*SD*)	95% CI	Mean (*SD*)	95% CI
Population means								
sf_Practice	0.009 (0.256)	−0.506, 0.523	−0.006 (0.244)	−0.502, 0.475	−0.049 (0.280)	−0.594, 0.504	−0.200 (0.314)	−0.824, 0.429
sf_Trial	−3.471 (1.065)	−5.671, −1.650	−3.001 (0.749)	−4.596, −1.669	−3.218 (1.051)	−5.513, −1.454	−2.733 (0.691)	−4.161, −1.545
Diffusion T1	3.795 (0.609)	2.702, 5.067	3.548 (0.445)	2.743, 4.467	3.748 (0.607)	2.688, 5.026	3.442 (0.466)	2.599, 4.432
Diffusion T2	3.523 (0.580)	2.538, 4.814	3.333 (0.431)	2.574, 4.247	3.512 (0.575)	2.487, 4.708	3.138 (0.417)	2.396, 4.009
Diffusion T3	2.781 (0.458)	2.025, 3.806	2.707 (0.352)	2.126, 3.439	2.801 (0.450)	1.993, 3.739	2.514 (0.348)	1.883, 3.244
Diffusion T4	5.499 (0.838)	4.089, 7.203	5.231 (0.650)	4.055, 6.515	4.816 (0.765)	3.449, 6.415	4.401 (0.582)	3.330, 5.583
Meas. error	0.299 (0.190)	0.080, 0.780	0.260 (0.173)	0.060, 0.713	0.233 (0.160)	0.054, 0.669	0.476 (0.923)	0.006, 2.934
Practice effect (PE)	0.399 (0.110)	0.184, 0.607	0.436 (0.117)	0.213, 0.665	0.388 (0.124)	0.145, 0.617	0.467 (0.123)	0.216, 0.714
Trial 1	6.087 (0.154)	5.781, 6.381	5.964 (0.163)	5.631, 6.263	5.939 (0.168)	5.611, 6.261	6.079 (0.161)	5.772, 6.389
Trial 2	8.782 (0.146)	8.514, 9.074	8.762 (0.149)	8.467, 9.046	8.663 (0.155)	8.366, 8.976	8.773 (0.141)	8.510, 9.056
Trial 3	9.837 (0.140)	9.578, 10.111	9.789 (0.141)	9.517, 10.055	9.733 (0.149)	9.428, 10.033	9.845 (0.136)	9.594, 10.114
Trial 4	8.478 (0.211)	8.054, 8.902	8.432 (0.212)	8.001, 8.839	8.297 (0.218)	7.877, 8.727	8.546 (0.199)	8.173, 8.925
Population correlations								
Trial 1–PE	0.013 (0.251)	−0.472, 0.474	−0.034 (0.224)	−0.461, 0.396	0.037 (0.315)	−0.552, 0.630	0.124 (0.331)	−0.511, 0.707
Trial 2–PE	0.186 (0.271)	−0.379, 0.681	0.102 (0.256)	−0.381, 0.588	0.199 (0.328)	−0.447, 0.766	0.293 (0.340)	−0.435, 0.811
Trial 3–PE	0.135 (0.280)	−0.463, 0.646	0.009 (0.260)	−0.495, 0.503	0.172 (0.344)	−0.505, 0.764	0.245 (0.363)	−0.502, 0.810
Trial 4–PE	0.178 (0.246)	−0.354, 0.627	0.048 (0.238)	−0.403, 0.487	0.194 (0.308)	−0.417, 0.727	0.249 (0.340)	−0.470, 0.785
Trial 2–Trial 1	0.897 (0.049)	0.783, 0.958	0.889 (0.058)	0.743, 0.967	0.901 (0.047)	0.778, 0.965	0.892 (0.057)	0.750, 0.961
Trial 3–Trial 1	0.816 (0.057)	0.686, 0.910	0.800 (0.077)	0.606, 0.911	0.824 (0.065)	0.659, 0.919	0.801 (0.076)	0.619, 0.911
Trial 4–Trial 1	0.654 (0.084)	0.476, 0.791	0.613 (0.113)	0.359, 0.798	0.634 (0.086)	0.447, 0.782	0.675 (0.099)	0.459, 0.833
Trial 3–Trial 2	0.927 (0.038)	0.832, 0.974	0.910 (0.051)	0.763, 0.972	0.925 (0.042)	0.827, 0.976	0.916 (0.045)	0.805, 0.974
Trial 4–Trial 2	0.817 (0.064)	0.674, 0.909	0.773 (0.089)	0.552, 0.901	0.799 (0.069)	0.620, 0.901	0.831 (0.076)	0.649, 0.931
Trial 4–Trial 3	0.867 (0.056)	0.724, 0.944	0.839 (0.077)	0.657, 0.942	0.843 (0.064)	0.685, 0.933	0.902 (0.055)	0.759, 0.973

95% CI, values are upper (2.5%) and lower (97.5%) bounds. sf_Practice, practice self-feedback; sf_Trial, trial self-feedback; Diffusion, standard deviation of diffusion processes for a given trial (e.g., T1 is Trial 1); Meas. error, measurement error; Trial represents manifest mean recall for each Trial, aggregated across occasions.

Age

The addition of years of age as a time independent predictor showed on average, older age was associated with worse performance on all four trials. However, this was qualified by interactions of mean trial performance with clinical diagnosis. Older age was associated with poorer performance on Trials 1, 2, and 3 among those with MCI diagnoses, but with trends toward better recall on trials 1 and 3 in those with DAT diagnoses (Figure 2). Moreover, the negative effect of dementia on PE gain per occasion was significant when accounting for age. A significant negative interaction of Age × MCI × Diffusion Trial 3 was due to older age attenuating the positive effects of MCI diagnosis on Trial 3 Diffusion. Here, whereas MCI diagnosis predicted higher levels of random variations in Trial 3 that benefited model prediction, this was limited by more advanced age.

Sex

Inclusion of participant sex in the model showed superior mean performance by women on all four recall trials. This was qualified by significant negative interactions of sex with diagnostic group predictors on Trial 1 for MCI and on all trials for DAT. As shown in the expectation plots (Figure 3), female sex was associated with lower performance in diagnosed DAT. In addition, significant positive interaction of sex with MCI on Trial 1 Diffusion, was due to higher Trial 1 Diffusion among women than men with MCI diagnoses. However, a significant negative interaction of sex with DAT on Trial 4 Diffusion showed lower predictive random changes in women than men with diagnoses of DAT on delayed recall trials.

Educational attainment

Greater educational attainment was marginally associated with higher performance on Trial 3 and 4 in the total sample. However, this was qualified by positive interactions between education and both MCI and DAT on Trial 4 only, where higher education predicted better delayed recall performance (Figure 4). More years of education also predicted lower measurement error in the MCI group, but higher measurement error in the DAT group.

TABLE 4 Effects of diagnostic groups in the total sample.

Interaction	Mean (*SD*)	95% CI
MCI × sf_Practice	0.063 (0.095)	−0.121, 0.248
MCI × sf_Trial	0.043 (0.291)	−0.548, 0.615
MCI × Diffusion T1	−0.477 (0.330)	−1.181, 0.136
MCI × Diffusion T2	0.284 (0.264)	−0.225, 0.800
MCI × Diffusion T3	0.466 (0.217)	0.052, 0.903*
MCI × Diffusion T4	0.012 (0.397)	−0.835, 0.743
MCI × Meas. error	0.007 (0.065)	−0.114, 0.148
MCI × Practice effect	−0.120 (0.133)	−0.372, 0.157
MCI × Trial 1	−1.423 (0.218)	−1.846, −0.996*
MCI × Trial 2	−1.769 (0.210)	−2.168, −1.356*
MCI × Trial 3	−1.778 (0.204)	−2.191, −1.382*
MCI × Trial 4	−2.923 (0.310)	−3.525, −2.350*
DAT × sf_Practice	−0.002 (0.103)	−0.198, 0.200
DAT × sf_Trial	0.182 (0.312)	−0.454, 0.802
DAT × Diffusion T1	−0.047 (0.404)	−0.835, 0.716
DAT × Diffusion T2	0.404 (0.323)	−0.219, 1.071
DAT × Diffusion T3	0.730 (0.307)	0.168, 1.379*
DAT × Diffusion T4	−0.948 (0.638)	−2.227, 0.215
DAT × Meas. error	−0.013 (0.071)	−0.160, 0.113
DAT × Practice effect	−0.435 (0.257)	−0.931, 0.068+
DAT × Trial 1	−2.328 (0.303)	−2.866, −1.741*
DAT × Trial 2	−3.807 (0.297)	−4.401, −3.211*
DAT × Trial 3	−4.409 (0.295)	−4.993, −3.839*
DAT × Trial 4	−6.345 (0.410)	−7.194, −5.554*

Values are mean with standard deviation in parentheses. 95% CI, values are upper (2.5%) and lower (97.5%) bounds. sf_Practice, practice self-feedback; sf_Trial, trial self-feedback; Diffusion, standard deviation of diffusion processes for a given trial (e.g., T1 is Trial 1); Meas. error, measurement error; Trial represents manifest mean recall for each Trial, aggregated across occasions. The asterisk * denotes significant effects; the + indicates nonsignificant trends.

Subsidiary models by diagnostic groups

In a series of models specific to each diagnostic group we also evaluated separate models with the time independent predictors age, sex, and educational attainment (Table 6). Complete details of all model outputs, including population means, population correlations and effects of time independent predictor are provided in Supplementary Material.

Cognitively unimpaired

The three models limited to the CU participants showed significant negative correlations between PE and mean recall performance on Trials 3 and 4 (Supplementary Table 2); those with better performance in the later and delayed recall trials had lower PE gain per occasion. Older age in CU participants was associated with higher Diffusion on Trials 3 and 4, and with lower overall mean performance on all trials (Table 6). In contrast, analysis of sex differences in the CU subsample showed men have higher Trial 3 diffusion and lower Trial 4 diffusion than women. Last, the education model showed higher educational attainment was associated with higher Trial 4 Diffusion, lower measurement error, and higher Trial 3 mean recall.

Mild cognitive impairment

Notably, the mean estimated PE gain per occasion parameter did not differ significantly from zero in the MCI subgroup analyses (Supplementary Table 3). In addition, MCI subgroup models did not show any significant correlations between mean Trial performance and PE gain. As with the CU analysis, older age in the MCI subgroup predicted lower Diffusion on Trials 3 and 4 and lower mean performance on Trials 2, 3, and 4. Modeling effects of sex in the MCI subgroups showed women to have better recall on Trials 3 and 4. Higher educational attainment in the MCI subgroup predicted better performance on Trials 1, 2, and 3, as well as lower overall PE.

Dementia

The DAT subgroup models showed significantly negative PE gain estimates (Supplementary Table 4). In addition, the subgroup models for Age and Sex both produced significant negative parameter estimates for the correlations between PE and mean performance on Trial 2 and Trial 4. Older age in the DAT subgroup predicted negative sf_Practice, but better mean performance on Trials 2 and 3. Sex differences were only manifest in mean level of Trial 3, where women performed worse than men. Higher educational attainment predicted higher Trial 4 Diffusion and trends for higher measurement error and lower PE, but no apparent effects on mean Trial performance.

Comparison of subgroup models

The individual models by clinical diagnosis demonstrated effects that were modified by the inclusion of specific covariates, as well as those that were consistent across subgroup sensitivity models. The sf_Practice parameter appeared sensitive to clinical diagnosis, with estimates that were more negative in the CU group and closer to zero in MCI; in contrast, estimated sf_Practice was positive in the DAT subgroup, and this was magnified by older age. In addition, PE gain was negatively correlated with Trial 3 and 4 recall performance for the CU subgroup; no correlations were significant between PE gain and performance in the MCI subgroup. The dementia subgroup showed higher PE was associated with lower recall performance only on Trials 2 and 4; however, this was only manifest in sensitivity models with Age and Sex and became non-significant when accounting for differences in Education. Similarly, the manifest means for all four recall Trials were consistently positively correlated across subgroup

TABLE 5 Significant and trending covariate effects and interactions with diagnostic groups in sensitivity models of age, sex and education.

Age	Mean (*SD*)	95% CI	Sex	Mean (*SD*)	95% CI	Education	Mean (*SD*	95% CI
Age × T1	−0.520 (0.152)	−0.820, −0.219	Sex × T1	0.357 (0.158)	0.038, 0.653	Educ. × T1	0.060 (0.059)	−0.053, 0.179
Age × T2	−0.255 (0.138)	−0.520, 0.007	Sex × T2	0.302 (0.136)	0.049, 0.566	Educ. × T2	0.077 (0.051)	−0.015, 0.186
Age × T3	−0.344 (0.130)	−0.614, −0.089	Sex × T3	0.280 (0.129)	0.026, 0.528	Educ. × T3	0.081 (0.048)	−0.008, 0.183
Age × T4	−0.521 (0.201)	−0.915, −0.125	Sex × T4	0.448 (0.207)	0.072, 0.841	Educ. × T4	−0.031 (0.069)	−0.167, 0.108
MCI × Diff. T1	−0.489 (0.305)	−1.129, 0.045	MCI × Diff. T1	−0.700 (0.339)	−1.445, −0.101	MCI × Diff. T1	−0.417 (0.317)	−1.052, 0.207
MCI × Diff. T2	0.389 (0.220)	−0.034, 0.778	MCI × Diff. T2	0.467 (0.220)	0.065, 0.931	MCI × Diff. T2	0.428 (0.203)	0.007, 0.818
DAT × Diff. T3	0.663 (0.281)	0.150, 1.241	DAT × Diff. T3	0.628 (0.301)	0.087, 1.283	DAT × Diff. T3	0.792 (0.271)	0.287, 1.324
DAT × Diff. T4	−1.355 (0.585)	−2.535, −0.269	DAT × Diff. T4	−0.949 (0.629)	−2.210, 0.253	DAT × Diff. T4	−1.435 (0.679)	−2.850, −0.149
DAT × PE	−0.514 (0.248)	−1.003, −0.020	DAT × PE	−0.463 (0.259)	−0.981, 0.039	DAT × PE	−0.472 (0.257)	−0.980, 0.004
Age × MCI × sf_Prac	0.020 (0.107)	−0.180, 0.231	Sex × MCI × sf_Prac	0.057 (0.101)	−0.149, 0.258	Educ. × MCI × sf_Prac	−0.097 (0.081)	−0.254, 0.063
Age × MCI × Diff. T1	0.050 (0.281)	−0.497, 0.576	Sex × MCI × Diff. T1	0.625 (0.312)	0.060, 1.273	Educ. × MCI × Diff. T1	0.178 (0.213)	−0.252, 0.554
Age × MCI × Diff. T3	−0.455 (0.219)	−0.893, −0.036	Sex × MCI × Diff. T3	0.126 (0.230)	−0.336, 0.581	Educ. × MCI × Diff. T3	0.045 (0.136)	−0.220, 0.308
Age × MCI × T1	−0.070 (0.204)	−0.441, 0.339	Sex × MCI × T1	−0.483 (0.227)	−0.925, −0.038	Educ. × MCI × T1	0.061 (0.089)	−0.114, 0.230
Age × MCI × T2	−0.385 (0.194)	−0.745, −0.019	Sex × MCI × T2	−0.167 (0.217)	−0.595, 0.256	Educ. × MCI × T2	0.040 (0.083)	−0.127, 0.197
Age × MCI × T3	−0.303 (0.186)	−0.673, 0.062	Sex × MCI × T3	−0.104 (0.207)	−0.503, 0.302	Educ. × MCI × T3	0.094 (0.081)	−0.070, 0.242
Age × MCI × T4	−0.660 (0.276)	−1.208, −0.149	Sex × MCI × T4	−0.208 (0.303)	−0.777, 0.390	Educ. × MCI × T4	0.213 (0.119)	−0.024, 0.447
Age × DAT × Diff. T4	−0.493 (0.576)	−1.623, 0.618	Sex × DAT × Diff. T4	−1.519 (0.628)	−2.800, −0.408	Educ. × DAT × Diff. T4	0.546 (0.433)	−0.328, 1.436
Age × DAT × ME	0.002 (0.058)	−0.107, 0.134	Sex × DAT × ME	0.005 (0.049)	−0.084, 0.110	Educ. × DAT × ME	0.129 (0.279)	−0.006, 0.985
Age × DAT × T1	0.549 (0.286)	−0.016, 1.130	Sex × DAT × T1	−0.646 (0.301)	−1.221, −0.048	Educ. × DAT × T1	−0.033 (0.126)	−0.276, 0.216
Age × DAT × T2	0.352 (0.265)	−0.147, 0.880	Sex × DAT × T2	−0.570 (0.283)	−1.118, −0.008	Educ. × DAT × T2	0.120 (0.119)	−0.111, 0.354
Age × DAT × T3	0.491 (0.258)	−0.010, 0.985	Sex × DAT × T3	−0.705 (0.266)	−1.240, −0.184	Educ. × DAT × T3	−0.014 (0.112)	−0.248, 0.202
Age × DAT × T4	−0.037 (0.355)	−0.729, 0.633	Sex × DAT × T4	−0.796 (0.376)	−1.499, −0.037	Educ. × DAT × T4	0.444 (0.156)	0.125, 0.752

Table depicts significant effects and interactions present in one or more of the three sensitivity models. Covariate effects that were not significant in any model are not shown. 95% CI, values are upper (2.5%) and lower (97.5%) bounds. MCI, diagnosis of amnestic or non-amnestic MCI; DAT, diagnosis of multi-domain amnestic dementia. sf_Prac, practice self-feedback; T, trial; for Diff. T1 is standard deviation of diffusion process for Trial 1; values of T1, T2, T3, T4 refer to manifest mean recall for each Trial, aggregated across occasions. PE, practice effect; ME, measurement error. Sex, men modeled as −1 and women as +1.

sensitivity models for CU and MCI subgroups, whereas the dementia subgroup showed more variable patterns across sensitivity models.

Comparison of covariate effects between subgroup and sensitivity models (Table 5) shows that older age was associated with lower mean performance on all trials for CU and MCI subgroups, and with higher performance on Trials 2 and 3 in the DAT subgroup (Table 6 and Supplementary Table 5). Moreover, whereas older age predicted higher Diffusion on Trials 3 and 4 in the CU subgroup, the opposite effect was manifest for the MCI group. In addition, older age only predicted more negative sf_Practice in those with diagnosed dementia. There were fewer effects of participant sex (Supplementary Table 6), although notably, while women in the MCI subgroup had higher Trial 3 performance, this was reversed in the DAT analysis. Higher educational attainment was associated with marginal benefits on mean performance on Trials 1–3 in the CU and MCI subgroups and with higher Trial 4 Diffusion in CU and DAT subgroups, but not MCI (Supplementary Table 7). Similarly, higher education was associated with lower PE only in the MCI subgroup.

Reparametrized to estimate within-occasion practice effects

The models reported in the present study focused on longitudinal practice effects. To address whether trial-by-trial improvements were also associated with clinical diagnosis we reparametrized the original diagnostic groups model to estimate relative within-occasion improvement. The new model estimated a parameter for Baseline performance as well as the deviations from Trial 1 for Trials 2, 3, and 4, rather than absolute performance, while otherwise maintaining the same model setup. Model results showed the population means (Supplementary Table 8) were consistent with the results from the original diagnostic groups model with one exception. The reparametrized model showed higher estimated level of Trial 3 Diffusion (mean = 4.873, sd = 0.384; 95% CI = 4.169–4.857) than the original (mean = 2.781, sd = 0.458; 95% CI = 2.025–3.806). Time independent predictor effects showed MCI and dementia diagnosis attenuated trial-by-trial improvements (Supplementary Table 9). Diagnosis did not interact with PE, measurement error, sf_Trial or sf_Practice, but both DAT and MCI diagnosis predicted significantly lower Trial

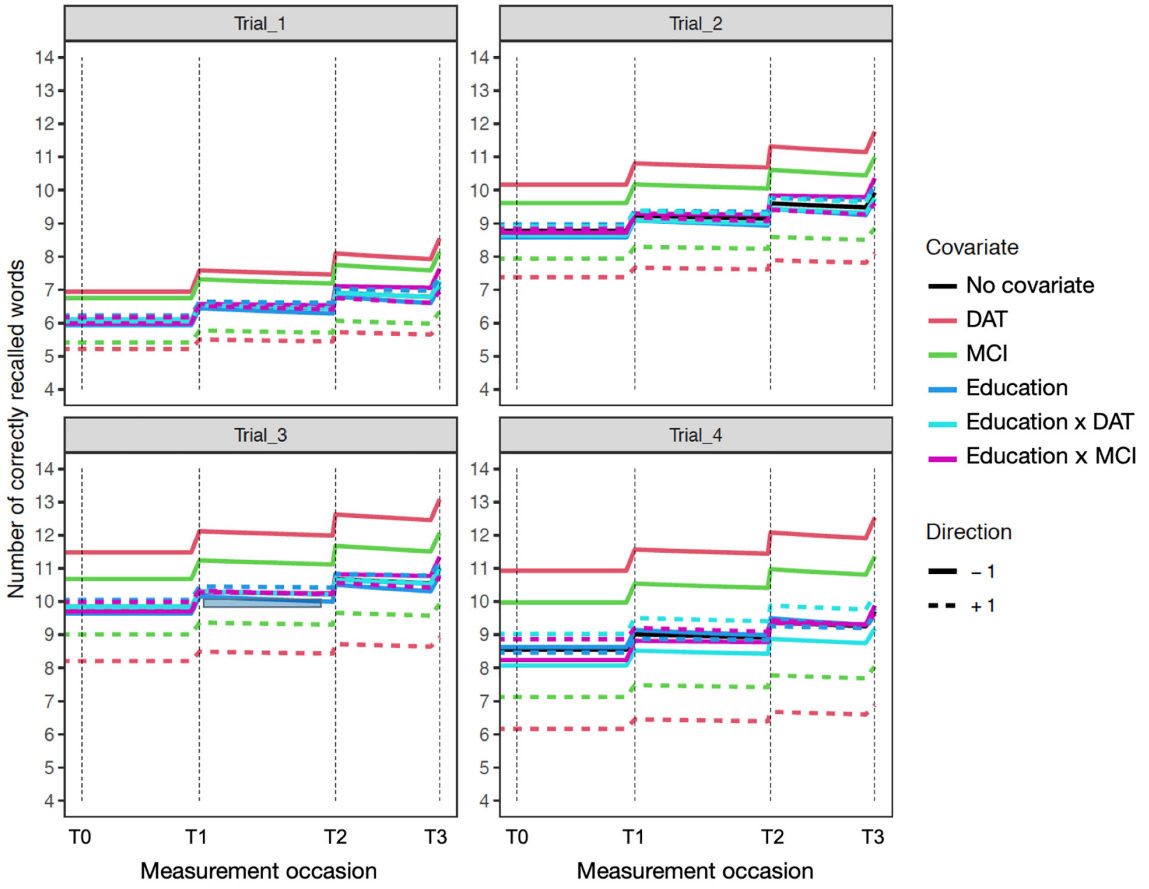


FIGURE 2
Education sensitivity model expectation plots for change in recall performance (black line) across four measurement occasions (e.g., T0: baseline, T1: 1 year). The solid and dashed colored lines depict effect modification by time independent predictors: DAT diagnosis (red lines), MCI diagnosis (green lines), Education (blue lines; higher: +1; lower: −1), and interactions of Education × DAT and Education × MCI. For covariate effects, higher (dashed) and lower (solid) levels of the covariate are shown to modify the level and expected slope. All covariate effects are in reference to 0 values of other covariates. The dashed red line reflects DAT, and the solid red line represents all other participants. Interaction effects only represent the total additive value of the interaction holding the main effects at zero. For example, in Trial_4 (lower right), the interaction of educational attainment and DAT shows higher education (+1) is associated with better performance in those with dementia diagnoses.

3 Diffusion. Inspection of the estimated population correlations between the trial-level deviations and parameter values for PE and Baseline performance showed higher baseline performance was associated with lower within-occasion improvement for Trial 2 and Trial 3 only (**Supplementary Table 9**). However, neither the trial-level deviations nor Baseline performance estimates were significantly correlated with PE. In addition, all three trial-level deviation parameters were positively correlated; greater improvement from Trial 1 tended to generalize across later trials.

Discussion

Dynamic modeling of PE from multi-occasion verbal learning data revealed multiple notable effects associated with clinical diagnoses of MCI and dementia. First, in accord with our initial hypotheses both manifest recall performance and overall PE varied as a function of diagnostic severity. In addition to diagnosis-specific variation in levels of mean trial performance, we observed a gradient of PE across the three diagnostic groups – from positive in CU participants to significantly negative in participants diagnosed with dementia. Whereas repeated performance conferred subsequent improvements in recall for unimpaired older adults, this was not consistently the case in those with diagnosed MCI; moreover, we observed ongoing decline in participants diagnosed with DAT, despite repeated testing, as evidenced by negatively estimated PE. Notably, modeling the four recall trials as individual processes permitted estimating mean performance separately from PE. Thus, mean trial performance is modeled as a stable, trait-like factor, whereas estimated PE served as the primary measure of

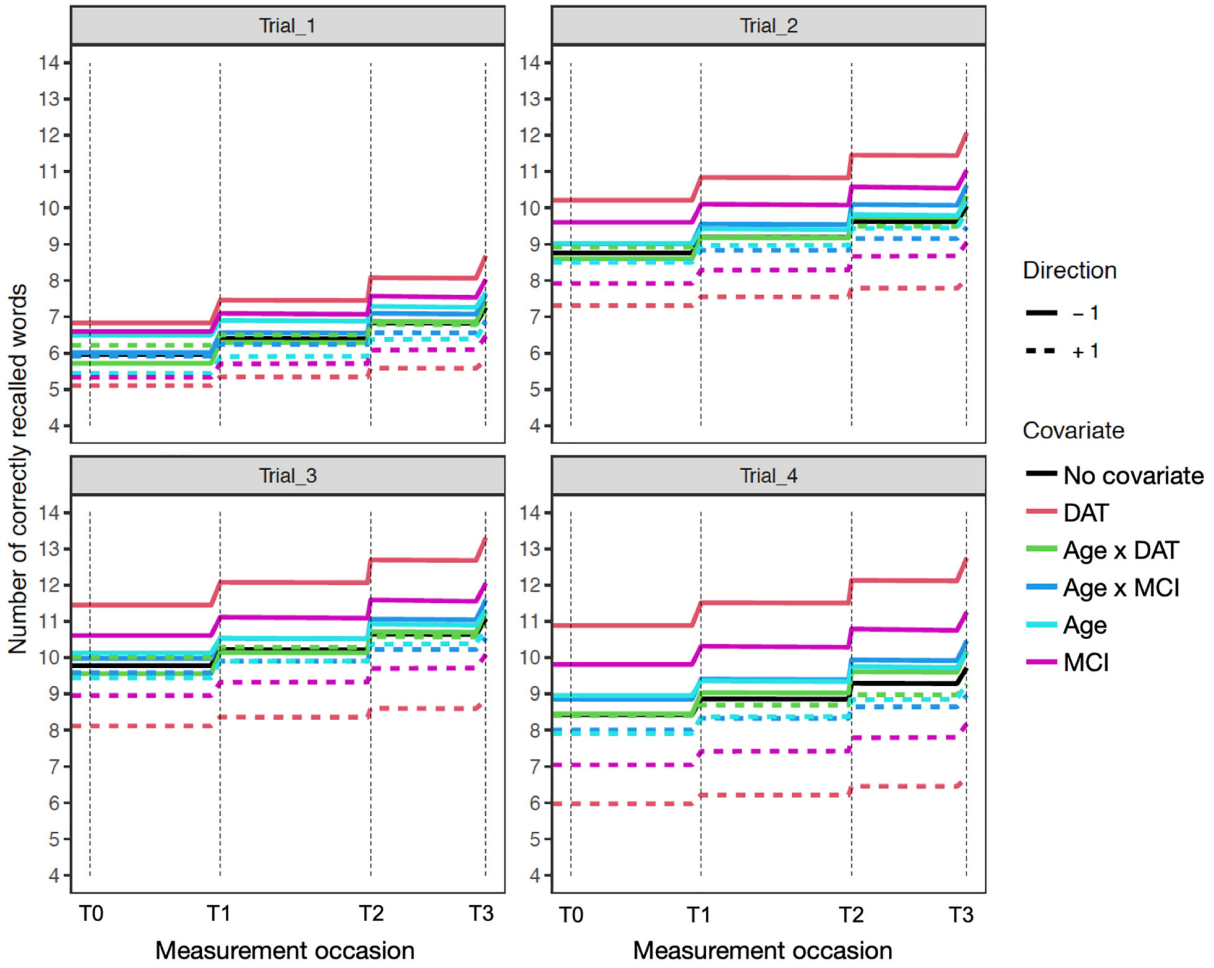


FIGURE 3
Age sensitivity model expectation plots for change in recall performance (black line) across four measurement occasions (e.g., T0: baseline, T1: 1 year). The solid and dashed colored lines depict effect modification by time independent predictors: DAT diagnosis (red lines), MCI diagnosis (magenta lines), Age (light blue lines), and interactions of Age × DAT and Age × MCI. For main effects, higher (dashed) and lower (solid) levels of the covariate modifies the level and expected patterns of change. All effects are in reference to 0 values on other covariates. The dashed red line reflects DAT, and the solid red line represents all other participants. Interaction effects only represent the total additive value of the interaction, when holding the main effects at zero.

change. This modeling perspective contrasts with most prior efforts that model PE as a linear change within or between occasions (Duff et al., 2007; Bender et al., 2013, 2020; Goldberg et al., 2015; Gavett et al., 2016). In addition, the PE parameter does not delineate between true decline and gains due to practice, as these are not considered separable processes in a dynamic system. The sensitivity of dynamic estimates of PE and performance to clinical diagnosis demonstrates the value of dynamic modeling in longitudinal clinical aging data.

Second, the present findings revealed previously unreported relationships between clinical diagnosis and dynamic process estimates. As with PE, the sf_Practice parameter followed a gradient of positive to negative values that corresponded with diagnostic severity. Practice self-feedback provides a non-linear measure of the extent level that practice (i.e., after completing all four trials for a given occasion) can boost or reduce estimated PE at the next occasion. The more positive estimates of sf_Practice in CU participants reflects an increase in practice-related gains on subsequent occasions. In contrast, both PE and sf_Practice were negatively estimated in participants with diagnosed dementia. Thus, while recall performance declined over time in these participants even with repeated testing (i.e., as indicated by negative PE estimates), dementia diagnosis was associated with less acceleration in decline. In other words, performance appears to stabilize at a lower level above floor in those diagnosed with dementia, despite both the absence of retest improvements and overall decline. In addition, both measurement error and diffusion processes (particularly on Trials 3 and 4), were sensitive to diagnostic group and other covariate effects. The standard

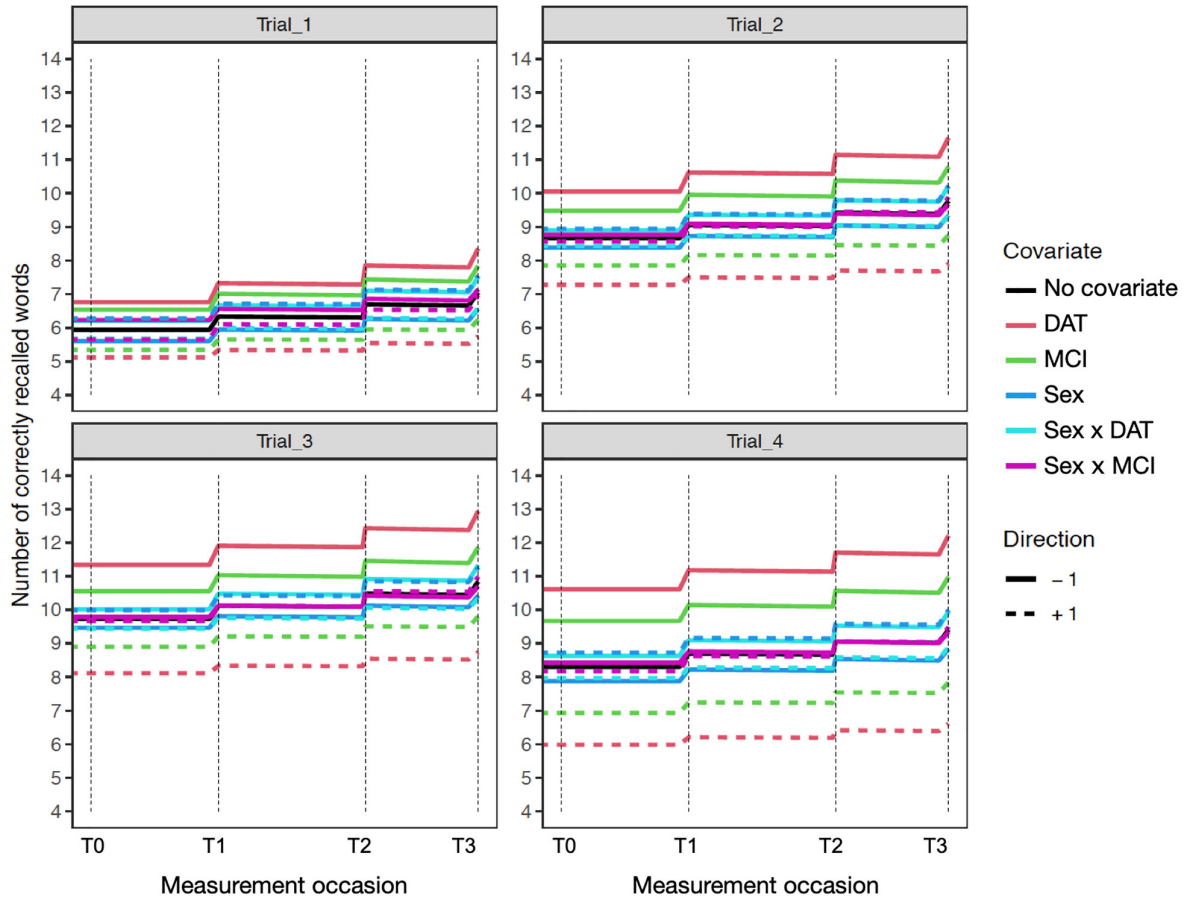


FIGURE 4
Expectation plots for change in recall performance (black line) across four measurement occasions (e.g., T0: baseline, T1: 1 year) in the sensitivity model of participant age. The solid and dashed colored lines depict effect modification by time independent predictors: DAT diagnosis (red lines), MCI diagnosis (green lines), Sex (blue lines; women: +1; men: −1), and interactions of Age × DAT and Age × MCI. For covariate effects, higher (dashed) and lower (solid) levels of the covariate are shown to modify the level and expected slope. All covariate effects are in reference to 0 values of other covariates.

deviation of the diffusion processes reflects unpredictable variations in trial performance that are nevertheless useful in predicting performance on other trials. In tasks like the HVLT, recall performance on later immediate recall trials necessarily includes savings from the preceding recall trials. The reported findings suggest that meaningful variations in Trial 3 performance may provide a uniquely sensitive marker of clinical cognitive impairment and dementia. Although the interpretation of such unpredictable variations is not clear, one possibility is Trial 3 diffusion processes may partly reflect impaired executive or amnestic encoding abilities. For example, reduced mental flexibility and working memory in MCI and dementia may produce more inconsistent recall performance across study occasions. Alternatively, higher Trial 3 diffusion processes may capture increasing reliance on list recency due to impaired short-term verbal encoding ability. Nevertheless, multiple cognitive processes are potentially implicated, which are likely to be further complicated by diagnosis and etiology. Therefore, additional work relating differences in non-linear PE estimates to more fine-grained neuropsychological performance is needed to clarify the cognitive processes responsible for variations in Trial 3 diffusion or other parameter estimates.

Third, evaluation of sensitivity and subgroup models revealed important sources of individual differences that modified multiple effects and qualified several interactions. For example, older age predicted worse mean recall on all Trials in the CU and diagnosed MCI subgroups, but better recall on trials 2 and 3 among those with dementia diagnoses (Table 6). This may suggest a survivor effect, as those who reach more advanced age before onset of dementia may maintain some residual abilities that enhance recall on these trials. Age also modified trial-specific diffusion processes for CU and MCI diagnosed participants, despite positive mean Diffusion estimates in both groups (Supplementary Tables 2, 3). Whereas older age predicted higher Diffusion on Trials 3 and 4 in the CU subgroup, the opposite effect was manifest for

TABLE 6 Significant and trending covariate effects of participant age, sex, and education on model parameters by subgroup.

	CU		MCI		DAT	
Interaction	**Mean (*sd*)**	**95% CI**	**Mean (*sd*)**	**95% CI**	**Mean (*sd*)**	**95% CI**
Age × sf_Practice	−0.002 (0.095)	−0.196, 0.175	0.044 (0.093)	−0.130, 0.217	−0.067 (0.033)	−0.130, −0.002*
Age × Diffusion T3	0.431 (0.210)	0.048, 0.909*	−0.568 (0.331)	−1.235, 0.085	−0.016 (0.360)	−0.750, 0.712
Age × Diffusion T4	0.699 (0.365)	0.034, 1.473*	−0.687 (0.409)	−1.499, 0.069	−0.060 (0.133)	−0.388, 0.138
Age × PE	0.080 (0.094)	−0.107, 0.260	−0.181 (0.131)	−0.435, 0.082	0.074 (0.201)	−0.330, 0.468
Age × Trial 1	−0.489 (0.140)	−0.763, −0.221*	−0.706 (0.139)	−0.980, −0.432*	0.163 (0.258)	−0.324, 0.627
Age × Trial 2	−0.305 (0.124)	−0.556, −0.067*	−0.660 (0.155)	−0.969, −0.375*	0.493 (0.203)	0.097, 0.893*
Age × Trial 3	−0.330 (0.104)	−0.532, −0.136*	−0.669 (0.145)	−0.972, −0.400*	0.473 (0.224)	0.047, 0.900*
Age × Trial 4	−0.482 (0.167)	−0.784, −0.175*	−1.200 (0.256)	−1.724, −0.703*	−0.090 (0.215)	−0.497, 0.313
Sex × Diffusion T3	−0.305 (0.194)	−0.693, 0.073	−0.279 (0.304)	−0.907, 0.307	−0.506 (0.348)	−1.284, 0.135
Sex × Diffusion T4	1.113 (0.322)	0.471, 1.737*	−0.019 (0.417)	−0.838, 0.785	−0.057 (0.133)	−0.403, 0.118
Education × Diffusion T4	0.753 (0.247)	0.336, 1.252*	0.305 (0.270)	−0.223, 0.852	0.450 (0.316)	0.035, 1.144+
Education × Meas. Error	−0.050 (0.010)	−0.070, −0.032*	−0.037 (0.050)	−0.123, 0.094	0.024 (0.083)	−0.033, 0.178
Education × PE	−0.018 (0.040)	−0.096, 0.062	−0.101 (0.058)	−0.217, 0.015+	−0.081 (0.080)	−0.240, 0.079
Education × Trial 1	0.035 (0.055)	−0.069, 0.146	0.107 (0.062)	−0.011, 0.227+	0.016 (0.112)	−0.205, 0.238
Education × Trial 2	0.063 (0.048)	−0.034, 0.159	0.095 (0.069)	−0.029, 0.233+	0.101 (0.094)	−0.088, 0.282
Education × Trial 3	0.072 (0.041)	−0.007, 0.153+	0.127 (0.065)	−0.002, 0.255+	0.074 (0.107)	−0.128, 0.277

Significant interactions denoted by asterisk (*). CU, cognitively unimpaired; MCI, diagnosis of amnestic or non-amnestic MCI; DAT, diagnosis of multi-domain amnestic dementia. Values are mean with standard deviation in parentheses. 95% CI, values are upper (2.5%) and lower (97.5%) bounds. sf_Practice, practice self-feedback; sf_Trial, trial self-feedback; Diffusion, standard deviation of diffusion processes for a given trial (e.g., T1 is Trial 1); Meas. error: measurement error; PE, practice effect gains; trial represents manifest mean recall for each Trial, aggregated across occasions. The + indicates nonsignificant trends.

the MCI group. This shows that in unimpaired adults older age enhances the generation of unpredictable but meaningful variation in performance but exerts the opposite effect in those diagnosed with MCI.

Greater education weakly predicted higher mean immediate recall abilities for CU and MCI. Higher educational attainment was also weakly associated with lower PE in the two subgroups with diagnoses of MCI or dementia (**Table 6**). Notably, estimated PE did not differ from zero in the MCI subgroup even though mean recall performance did not show a ceiling effect; furthermore, PE and recall were not significantly associated in this group. Thus, greater educational attainment in the presence of manifest cognitive impairment may predict greater loss of neurocognitive plasticity necessary to benefit from repetition. Critically, this finding should be interpreted in the context of recent reports showing education does not appear neuroprotective or to confer resilience to cognitive decline or neurodegeneration. Rather, more years of early-life education may increase premorbid level of ability and positively offset trajectories of decline (Wilson et al., 2019; Lovden et al., 2020; Nyberg et al., 2021). However, for those whose neurocognitive abilities have reached a functional threshold for impairment, higher education may be associated with accelerated declines. Here, PE appears to be a marker of such accelerated functional declines. Similarly, the advantage of female sex on tests of verbal memory (Herlitz et al., 1997; Herlitz and Rehnman, 2008; Bender et al., 2010) was largely negligible, with one notable exception. Whereas women in the MCI diagnosis subgroup had better mean recall on Trial 3, this was reversed in the more impaired participants with dementia diagnoses. As with education, it is possible that this reflects a positive offset in trajectories of decline due to higher premorbid level of verbal memory abilities, resulting in steeper declines following onset of dementia.

Under the present dynamic modeling framework, performance on each occasion reflects ongoing processes that are inherently altered by prior testing exposure or experience. Here, PE reflects total intra-person change as the combination of maintenance or decline in addition to contributions of prior experience. Therefore, while the variable gains or losses following practice are not clearly dissociable from ongoing declines, separating level from change across trials captures multiple behavioral dimensions relevant to clinical diagnosis. For example, we note that the relationship between better recall performance and lower PE was only observed in the participants with CU or dementia diagnoses, but not in those with MCI. Unimpaired participants performing closer to ceiling had less room to improve and were more likely to show reduced subsequent gains. In contrast, the negative estimates of PE in the subgroup with dementia diagnoses captures longitudinal declines – those with higher overall performance also had the furthest to decline. However, the disconnection between PE and level of performance in MCI suggests these two dimensions may provide unique diagnostic or prognostic information. This aligns with prior findings showing PE differences are a meaningful indicator of progressive decline in older adults with

MCI diagnoses who exhibit low-to-moderate levels of recall performance (Duff et al., 2007; Rabin et al., 2009; Hassenstab et al., 2015; Gavett et al., 2016).

The present findings point to greater inconsistencies in responses, across trials and occasions as additional markers of cognitive impairment and dementia. We found that mean recall performance was consistently correlated across trials in the CU and MCI subgroups, but not in those diagnosed with dementia. Furthermore, correlations among Trials for the dementia subgroup showed more variable patterns across sensitivity models. In addition to declines in PE and mean recall performance, it is possible that loss of neurocognitive plasticity may also manifest as less consistent responses. Although the models were specified to focus primarily on longitudinal practice effects, such inconsistency may reflect reduced within-occasion improvement across trials. We also observed MCI and DAT diagnoses attenuated trial-by-trial improvements in the reparametrized model. Similarly, reduced short-term PE has previously been related to differences in clinical diagnosis, cognitive function, and brain structure (Duff et al., 2007, 2012; Fernandez-Ballesteros et al., 2012; Bender et al., 2020). These findings support the view that dynamic estimates of PE within and across occasions provides meaningful proxies for cognitive plasticity associated with advanced age or pathology (Baltes and Raykov, 1996; Yang, 2011). Further work is needed to identify which aspects of PE provide the most sensitive behavioral markers of ongoing declines.

Limitations and future directions

The present findings provide important evidence regarding the value of dynamic modeling approaches in estimating longitudinal change in performance as a function of PE. The limitations in the present study methods and findings must be acknowledged, while also highlighting corresponding opportunities for further inquiry. The model treated clinical diagnosis as a time independent predictor, but this did not accurately represent the diagnostic variability manifest in 11% of the study sample across study occasions. Six participants with initial diagnoses of amnestic MCI converted to DAT, and an additional six participants originally characterized as CU later received diagnoses of MCI. In addition, 15 participants with baseline MCI diagnoses were characterized as CU at their most recent visit. Although the modeling approach used here did not attempt to account for such variation in diagnosis, further work is needed to evaluate dynamic modeling for more transient changes in cognitive status. One alteration from the present approach could be to model diagnosis as a time varying measure, provided sufficient variation is present. Similarly, data sampled more intensively or with more variable timing would also make better use of capacity for modeling time in *ctsem*. While it is possible that accounting for such variation in assessed cognitive status may affect the results, future work should examine how intra-individual variation in clinical diagnosis is manifest in HVLT recall performance and PE. Similarly, the dementia subgroup only included participants with Alzheimer's (including mixed dementia) and including participants with other forms of dementia associated with other neurodegenerative diseases such as Lewy bodies, fronto-temporal dementia, or posterior cortical atrophy may demonstrate further sensitivity of PE and dynamic performance estimates to underlying pathologies. Furthermore, the same stimulus lists were presented on each occasion in the present study; future work should compare the effects of repeated vs. non-repeated content.

In addition, the available data for participants with dementia diagnoses was limited to three observations, although these were distributed across the actual occasions of assessment. While most statistical methods typically focus only on observed data, prior findings show patients with moderate Alzheimer's dementia are more prone to non-response (Feng et al., 2020; Wang et al., 2021). The HVLT is a challenging task for patients with mild to moderate dementia and patients may become quickly discouraged. The modeling of non-ignorable missingness for statistical inference is a daunting task in practice owing to its unknown nature and non-identifiable model parameters. Although challenging, additional research is needed on further implementation of methods for modeling informative missingness in the context of estimating PE in a Bayesian structural equation modeling framework.

The reported findings suggest that meaningful variations in Trial 3 performance may provide a uniquely sensitive marker of clinical cognitive impairment and dementia. This may show that certain trials are more important in HVLT performance and PE, which could be useful in clinical applications. More work is needed to shed light on differences in individual trials and their potential utility in clinical applications. However, this would require substantially more individual data to generate population-based normative estimates for direct comparison with individual patient cases. Similarly, for other potential applications of these methods, such as power estimation for dementia prevention trials, a larger number of normative data would help reduce uncertainty in parameter estimates (i.e., shrink confidence intervals) for such complex dynamic models. Notably, prevention trials tend to have rigorously standardized schedules of assessment, while *ctsem* benefits from more variable timing across assessments in order to reduce uncertainty. Thus, clinical trials may benefit from increased flexibility in timing to better leverage dynamic modeling approaches. New methods for intensive behavior sampling using smartphones provide a clear opportunity to bridge this divide, as they allow for considerably more dense measurement and greater variability in timing. Future work should evaluate dynamic models of PE in large, normative data sets from acquired with such methods.

In addition, the present study only evaluated HVLT task data with 12 words per recall trial; however, the use of longer lists of 15 or 16 words in other verbal learning tasks could

conceivably modify effects where CU participants performed close to ceiling. Future work should evaluate the effect of differences in cognitive load as a function of varying lengths of study lists. Similarly, the verbal nature of the data may confound lexical fluency with memory and PE; future investigations applying dynamic modeling approaches to estimate PE in non-verbal tasks and response times.

The present study generated far more testable hypotheses than it directly addressed. Nevertheless, the findings reported here demonstrate the expanded potential for evaluating new measures of performance affected by aging, neurodegeneration, or clinical diagnosis afforded by modeling non-linear dynamic processes.

Conclusion

The present findings highlight the sensitivity of dynamically modeled estimates of PE and verbal recall to diagnosed MCI and dementia. Modeling PE as the primary measure of change of showed PE gains and non-linear practice self-feedback, as well as mean level of recall performance are sensitive to severity of cognitive impairment and clinical dementia diagnosis. Moreover, applying dynamic modeling to longitudinal verbal learning data captures new behavioral dimensions reflecting intra-individual variations that are sensitive to cognitive impairment and dementia. Dynamic modeling using the ctsem framework provides a new perspective for modeling longitudinal changes in performance due to aging and dementia.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: the dataset analyzed for this study was provided as a Limited Dataset under a signed Data Use Agreement. As such it is not publicly available for analysis. This dataset will be made available to researchers only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to attempt to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed. Requests to access these datasets should be directed to Arijit K. Bhaumik, Research Administrator, arijit@med.umich.edu.

Ethics statement

The studies involving human participants were reviewed and approved by Michigan State University Human Research Protection Program. The patients/participants provided their written informed consent to participate in this study.

Author contributions

BH oversaw data collection. CD developed the model in *ctsem* in consultation with AB. CD developed the code for modeling and data visualization. AG implemented and ran all models and output figures. AB interpreted model results in consultation with CD and AG. AB drafted the manuscript with assistance from MM. AG, BH, and CD contributed essential revisions. All authors approved the final version of the manuscript for submission.

Funding

This project was partially supported by the NIH/NIA funded Michigan Alzheimer's Disease Research Center (5P30AG053760 and 1P30AG072031) and the University of Michigan Claude D. Pepper Older Americans Independence Center (1P30AG024824). Support to BH from NIA R35AG072262 (effort) is also acknowledged.

Acknowledgments

We extend their gratitude to the patients, caregivers, and families whose participation in research studies at the Michigan Alzheimer's Disease Research Center makes this work possible.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.911559/full#supplementary-material

References

Baltes, M. M., and Raykov, T. (1996). Prospective validity of cognitive plasticity in the diagnosis of mental status: a structural equation model. *Neuropsychology* 10:549.

Bender, A. R., and Raz, N. (2012). Age-related differences in recognition memory for items and associations: contribution of individual differences in working memory and metamemory. *Psychol. Aging* 27, 691–700. doi: 10.1037/a0026714

Bender, A. R., Brandmaier, A. M., Duzel, S., Keresztes, A., Pasternak, O., Lindenberger, U., et al. (2020). Hippocampal Subfields and Limbic White Matter Jointly Predict Learning Rate in Older Adults. *Cereb. Cortex* 30, 2465–2477. doi: 10.1093/cercor/bhz252

Bender, A. R., Daugherty, A. M., and Raz, N. (2013). Vascular risk moderates associations between hippocampal subfield volumes and memory. *J. Cogn. Neurosci.* 25, 1851–1862. doi: 10.1162/jocn_a_00435

Bender, A. R., Naveh-Benjamin, M., and Raz, N. (2010). Associative deficit in recognition memory in a lifespan sample of healthy adults. *Psychol. Aging* 25, 940–948. doi: 10.1037/a0020595

Blasi, S., Zehnder, A. E., Berres, M., Taylor, K. I., Spiegel, R., and Monsch, A. U. (2009). Norms for change in episodic memory as a prerequisite for the diagnosis of mild cognitive impairment (MCI). *Neuropsychology* 23, 189–200. doi: 10.1037/a0014079

Brandt, J. (1991). The Hopkins Verbal Learning Test: development of a new memory test with six equivalent forms. *Clin. Neuropsychol.* 5, 125–142.

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Campos-Magdaleno, M., Facal, D., Lojo-Seoane, C., Pereiro, A. X., and Juncos-Rabadan, O. (2017). Longitudinal Assessment of Verbal Learning and Memory in Amnestic Mild Cognitive Impairment: Practice Effects and Meaningful Changes. *Front. Psychol.* 8:1231. doi: 10.3389/fpsyg.2017.01231

Driver, C. C., and Voelkle, M. C. (2021). "Hierarchical continuous time modeling," in *The Handbook of Personality Dynamics and Processes*, ed. J. Rauthmann (Amsterdam: Elsevier), 887–908.

Driver, C. C., Oud, J. H., and Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *J. Stat. Softw.* 77, 1–35.

Duff, K., Beglinger, L. J., Schultz, S. K., Moser, D. J., McCaffrey, R. J., Haase, R. F., et al. (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch. Clin. Neuropsychol.* 22, 15–24. doi: 10.1016/j.acn.2006.08.013

Duff, K., Callister, C., Dennett, K., and Tometich, D. (2012). Practice effects: a unique cognitive variable. *Clin. Neuropsychol.* 26, 1117–1127. doi: 10.1080/13854046.2012.722685

Duff, K., Suhrie, K. R., Dalley, B. C. A., Anderson, J. S., and Hoffman, J. M. (2019). External validation of change formulae in neuropsychology with neuroimaging biomarkers: a methodological recommendation and preliminary clinical data. *Clin. Neuropsychol.* 33, 478–489. doi: 10.1080/13854046.2018.1484518

Duff, K., Westervelt, H. J., McCaffrey, R. J., and Haase, R. F. (2001). Practice effects, test–retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. *Arch. Clin. Neuropsychol.* 16, 461–476.

Feng, X., Li, T., Song, X., and Zhu, H. (2020). Bayesian Scalar on Image Regression With Nonignorable Nonresponse. *J. Am. Stat. Assoc.* 115, 1574–1597. doi: 10.1080/01621459.2019.1686391

Fernandez-Ballesteros, R., Botella, J., Zamarron, M. D., Molina, M. A., Cabras, E., Schettini, R., et al. (2012). Cognitive plasticity in normal and pathological aging. *Clin. Interv. Aging* 7, 15–25. doi: 10.2147/CIA.S27008

Gavett, B. E., Gurnani, A. S., Saurman, J. L., Chapman, K. R., Steinberg, E. G., Martin, B., et al. (2016). Practice Effects on Story Memory and List Learning Tests in the Neuropsychological Assessment of Older Adults. *PLoS One* 11:e0164492. doi: 10.1371/journal.pone.0164492

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., and Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement.* 1, 103–111. doi: 10.1016/j.dadm.2014.11.003

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Hawkins, K. A., Dean, D., and Pearlson, G. D. (2004). Alternative forms of the Rey Auditory Verbal Learning Test: a review. *Behav. Neurol.* 15, 99–107. doi: 10.1155/2004/940191

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., and 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. *Clin. Neuropsychol.* 24, 1267–1278. doi: 10.1080/13854046.2010.526785

Herlitz, A., and Rehnman, J. (2008). Sex differences in episodic memory. *Curr. Dir. Psychol. Sci.* 17:52.

Herlitz, A., Nilsson, L. G., and Bäckman, L. (1997). Gender differences in episodic memory. *Mem. Cogn.* 25, 801–811.

Hertzog, C., and Dunlosky, J. (2004). Aging, metacognition, and cognitive control. *Psychol. Learn. Motiv.* 45, 215–251.

Hoffman, L., Hofer, S. M., and Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: a simulation study. *Psychol. Aging* 26, 778–791. doi: 10.1037/a0023910

Karpicke, J. D., Lehman, M., and Aue, W. R. (2014). "Retrieval-Based Learning: an episodic context account," in *The Psychology of Learning and Motivation*, ed. B. H. Ross (San Diego, CA: Elsevier Academic Press), 237–284.

Lonie, J. A., Parra-Rodriguez, M. A., Tierney, K. M., Herrmann, L. L., Donaghey, C., O'Carroll, R. E., et al. (2010). Predicting outcome in mild cognitive impairment: 4-year follow-up study. *Br. J. Psychiatry* 197, 135–140. doi: 10.1192/bjp.bp.110.077958

Lovden, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., and Tucker-Drob, E. M. (2020). Education and Cognitive Functioning Across the Life Span. *Psychol. Sci. Public Interest* 21, 6–41. doi: 10.1177/1529100620920576

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

Nyberg, L., Magnussen, F., Lundquist, A., Baare, W., Bartres-Faz, D., Bertram, L., et al. (2021). Educational attainment does not influence brain aging. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2101644118. doi: 10.1073/pnas.2101644118

Rabbitt, P., Diggle, P., Smith, D., Holland, F., and Mc Innes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia* 39, 532–543. doi: 10.1016/s0028-3932(00)00099-3

Rabin, L. A., Pare, N., Saykin, A. J., Brown, M. J., Wishart, H. A., Flashman, L. A., et al. (2009). Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer's disease. *Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn.* 16, 357–376. doi: 10.1080/13825580902825220

Raykov, T., Baltes, M. M., Neher, K. M., and Sowarka, D. (2002). A comparative study of two psychometric approaches to detect risk status for dementia. *Gerontology* 48, 185–193.

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology* 24, 563–572. doi: 10.1037/a0019026

Salthouse, T. A., and Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology* 22, 800–811. doi: 10.1037/a0013091

Salthouse, T. A., Schroeder, D. H., and Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Dev. Psychol.* 40, 813–822. doi: 10.1037/0012-1649.40.5.813

Sanchez-Benavides, G., Gispert, J. D., Fauria, K., Molinuevo, J. L., and Gramunt, N. (2016). Modeling practice effects in healthy middle-aged participants of the Alzheimer and Families parent cohort. *Alzheimers Dement.* 4, 149–158. doi: 10.1016/j.dadm.2016.07.001

Summers, M. J., and Saunders, N. L. (2012). Neuropsychological measures predict decline to Alzheimer's dementia from mild cognitive impairment. *Neuropsychology* 26, 498–508. doi: 10.1037/a0028576

Voelkle, M. C., and Oud, J. H. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating

processes. *Br. J. Math. Stat. Psychol.* 66, 103–126. doi: 10.1111/j.2044-8317.2012.02043.x

Voelkle, M. C., Oud, J. H., Davidov, E., and Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychol. Methods* 17:176. doi: 10.1037/a0027543

Wang, X., Song, X., and Zhu, H. (2021). Bayesian latent factor on image regression with nonignorable missing data. *Stat. Med.* 40, 920–932. doi: 10.1002/sim.8810

Wilson, B. A., Watson, P. C., Baddeley, A. D., Emslie, H., and Evans, J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *J. Int. Neuropsychol. Soc.* 6, 469–479.

Wilson, R. S., Yu, L., Lamar, M., Schneider, J. A., Boyle, P. A., and Bennett, D. A. (2019). Education and cognitive reserve in old age. *Neurology* 92, e1041–e1050. doi: 10.1212/WNL.0000000000007036

Yang, L. (2011). Practice-oriented retest learning as the basic form of cognitive plasticity of the aging brain. *J. Aging Res.* 2011:407074. doi: 10.4061/2011/407074

Yang, L., and Krampe, R. T. (2009). Long-term maintenance of retest learning in young old and oldest old adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 64, 608–611. doi: 10.1093/geronb/gbp063

Zimprich, D., Rast, P., and Martin, M. (2008). "Individual differences in verbal learning in old age," in *Handbook of Cognitive Aging: Interdisciplinary Perspectives*, eds S. M. Hofer and D. F. Alwin (Thousand Oaks: Sage Publications, Inc), 224–243.

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 23 September 2022
DOI 10.3389/fnagi.2022.936528

OPEN ACCESS

EDITED BY
Holly Jeanne Bowen,
Southern Methodist University,
United States

REVIEWED BY
Leslie Susan Gaynor,
University of California, San Francisco,
United States
Nelson Roque,
University of Central Florida,
United States

*CORRESPONDENCE
Chandramallika Basak
cbasak@utdallas.edu

SPECIALTY SECTION
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 10 May 2022
ACCEPTED 29 August 2022
PUBLISHED 23 September 2022

CITATION
Smith ET, Skolasinska P, Qin S, Sun A, Fishwick P, Park DC and Basak C (2022) Cognitive and structural predictors of novel task learning, and contextual predictors of time series of daily task performance during the learning period.
Front. Aging Neurosci. 14:936528.
doi: 10.3389/fnagi.2022.936528

Cognitive and structural predictors of novel task learning, and contextual predictors of time series of daily task performance during the learning period

Evan T. Smith[1,2], Paulina Skolasinska[1,2], Shuo Qin[1], Andrew Sun[1], Paul Fishwick[3], Denise C. Park[1,2] and Chandramallika Basak[1,2]*

[1]Center for Vital Longevity, University of Texas at Dallas, Dallas, TX, United States, [2]Department of Psychology, University of Texas at Dallas, Richardson, TX, United States, [3]School of Arts and Technology, University of Texas at Dallas, Richardson, TX, United States

Investigation into methods of addressing cognitive loss exhibited later in life is of paramount importance to the field of cognitive aging. The field continues to make significant strides in designing efficacious cognitive interventions to mitigate cognitive decline, and the very act of learning a demanding task has been implicated as a potential mechanism of augmenting cognition in both the field of cognitive intervention and studies of cognitive reserve. The present study examines individual-level predictors of complex skill learning and day-to-day performance on a gamified working memory updating task, the BirdWatch Game, intended for use as a cognitive intervention tool in older adults. A measure of verbal episodic memory and the volume of a brain region involved in verbal working memory and cognitive control (the left inferior frontal gyrus) were identified as predictors of learning rates on the BirdWatch Game. These two neuro-cognitive measures were more predictive of learning when considered in conjunction than when considered separately, indicating a complementary effect. Additionally, auto-regressive time series forecasting analyses were able to identify meaningful daily predictors (that is, mood, stress, busyness, and hours of sleep) of performance-over-time on the BirdWatch Game in 50% of cases, with the specific pattern of contextual influences on performance being highly idiosyncratic between participants. These results highlight the specific contribution of language processing and cognitive control abilities to the learning of the novel task examined in this study, as well as the variability of subject-level influences on task performance during task learning.

KEYWORDS

game learning, cognitive training, time-series analysis, aging, gray matter volume, game intervention design

Introduction

Investigation of factors that influence successful learning has a long history in the psychological sciences. Aside from obvious importance to the fields of learning and skill development, the question of what factors influence individual learning rates is also of central importance to the field of cognitive training in normal aging. Several investigations of cognitive training have found that learning outcomes during the training period directly relate to training outcomes in terms of transfer to unrelated or "far" cognitive measures (Basak et al., 2008; Bürki et al., 2014; Basak and O'Connell, 2016). Based on their findings, Bürki et al. (2014) concluded that an understanding of the individual difference factors that influence the learning of the training task is a critical step in the development of efficacious cognitive intervention, and other researchers have expressed a similar position (Taatgen, 2013; Gathercole et al., 2019). Past research in this domain has revealed several cognitive and brain structure factors which appear to predict success in novel complex task learning in older adults (Erickson et al., 2010; Basak et al., 2011; Ray et al., 2017; Smith et al., 2020).

Both cognitive and structural predictors of learning novel, computerized tasks have been identified by past research. Ray et al. (2017) reported that measures of working memory were predictive of learning rates of two novel video games in a lifespan sample. This finding was later replicated by Smith et al. (2020). In addition to working memory, Ray et al. (2017) found a measure of perceptual discrimination (cued discrimination task; Posner, 1980) to be predictive of learning for the strategy game that relied more on working memory and cognitive control than the action game. In terms of structural predictors, in younger adults, Erickson et al. (2010) demonstrated that individual differences in the gray matter volume (GMV) of the striatum predicted learning outcomes on a lab-developed game-like computer task designed to stress working memory, cognitive control, and response time. In older adults, Basak et al. (2011) identified a number of predominantly left fronto-parietal gray matter regions (including left medial frontal gyrus, left dorsolateral prefrontal cortex, anterior cingulate cortex, and left postcentral gyrus) and cerebellum, whose volumes predicted learning of a commercial real-time strategy video game, which had shown transfer to laboratory-based measures of cognitive control, working memory, and reasoning (Basak et al., 2008). White matter correlates of novel computer task learning have also been identified: Ray et al. (2017) identified two discreet white matter microstructures (left cingulum-hippocampus and right fornix-stria terminalis), the integrity of which predicted the learning rate on two commercial video games. Importantly, left cingulum-hippocampus integrity predicted learning in the strategy game in both young and old adults. In sum, left fronto-parietal gray matter volumes and structural connectivity between the hippocampus and frontal cortex have been predictive of novel strategy game learning in older adults.

Another factor that may strongly contribute to individual differences in task learning, especially in older adults, is cognitive reserve. Cognitive reserve is known to be predictive of performance on episodic and working memory tasks, executive function, speed of processing, and general cognition (Opdebeeck et al., 2016). Considering that all of these factors are likely invoked in the learning of a complex, novel task, such as those used in cognitive training interventions (Gathercole et al., 2019), and the known relationship between cognitive reserve and retained cognitive function in later life (Park et al., 2014; Bak et al., 2016; Ward et al., 2020), an investigation of how cognitive reserve interacts with novel task learning is similarly warranted.

As this body of work demonstrates, the field is continuously making strides in identifying individual difference factors that influence the learning of novel tasks. However, if our stated goal is to apply this knowledge to develop efficacious cognitive interventions for at-risk groups, particularly the elderly, the above-summarized research exhibits some limitations. First, most of the studies cited above used a young adult (Erickson et al., 2010) or lifespan sample (Ray et al., 2017; Smith et al., 2020), which limits the conclusions we can draw with regard to our target population, that is, older adults aged 65 years and above. Second, all but one of the above-cited studies (Basak et al., 2011 being the exception) utilized short-term learning periods of 2.5 h or less, which therefore limits any conclusions we can draw from this research to this early period of task learning. As most reported cognitive interventions in older adults are of a substantially greater length (for a meta-analysis, see Basak et al., 2020), an examination of how such predictors affect learning at a later training phase is warranted. Third, the act of task learning requires consistent invocation of episodic memory, working memory, and cognitive control (Taatgen, 2013), and these capacities are susceptible to a wide range of cognitive and psychosocial contextual factors (Stawski et al., 2011). Considering this, it is likely that such factors have a downstream influence on the task learning process itself, which may contribute to the large individual differences in patterns of task learning that have been observed (Bürki et al., 2014), but examinations of such contextual effects on performance during training tasks are lacking.

Based on the findings and limitations of the above-summarized research, the present study was designed to further examine cognitive and structural correlates of learning on a working memory training task, as well as daily contextual factors which may influence training task performance during the training period. Reasoning and episodic memory were selected as cognitive predictors in order to expand on the past research which has already established working memory ability as a correlate of task learning (Ray et al., 2017; Smith et al., 2020). To evaluate the cognitive and structural correlates and daily contextual factors of learning on a training task, we

used data from a recently completed clinical trial in healthy aging (registered at ClinicalTrials.gov as NCT03988829), where variations of a PI-developed working memory training game ("BirdWatch Game") were used as interventions. For the present study, we focused on the BirdWatch Game and baseline measures of hypothesized cognition and gray matter volume correlates of learning of that game. If episodic memory and reasoning interact with BirdWatch Game learning as working memory has been demonstrated to with other computerized task learning, we would expect participants with greater pre-training ability on those constructs to demonstrate more rapid learning of the BirdWatch Game, and potentially higher maximum attainment. Additionally, because the BirdWatch Game itself is a working memory updating training paradigm, initial performance on the BirdWatch Game can be interpreted as the baseline working memory ability (both capacity and updating) of participants in this study. By that conceptualization, we predict that individuals with greater initial performance on the BirdWatch Game will show more rapid learning of the task, in line with past research (Ray et al., 2017). We hypothesize that cognitive reserve will demonstrate a similar relationship to task learning as the other examined cognitive constructs, considering past research which has observed a correlation between cognitive reserve and initial task learning (Lojo-Seoane et al., 2020). Alternatively, lower cognitive construct/reserve measures prior to training may relate instead to greater improvement on the trained task due to lower initial performance, as similar results have been observed in some past cognitive training studies (López-Higes et al., 2018). We expect this alternate hypothesis to be supported by greater progress in late learning specifically, if indeed it is supported, considering the past evidence that relatively lower cognitive ability/reserve results in slow initial learning (Ray et al., 2017; Lojo-Seoane et al., 2020).

A recent meta-analysis on cognitive interventions across both healthy aging and older adults with mild cognitive impairments (Basak et al., 2020), which included 214 cognitive training studies, found that the immediate cognitive gains in the cognitive training group is significantly more than the control group (net gain effect size $= 0.28$, $p < 0.001$). Importantly, the most effective intervention that resulted in the largest effects of near and far transfer trained either executive functions or working memory. The PI and her team designed a computerized cognitive training intervention, the BirdWatch Game, based on the Theory of Working Memory Adaptability (Basak and O'Connell, 2016), which predicts that high cognitive control demands from unpredictable probe-cues during working memory updating engender greater far transfer than predictable probe-cues in healthy aging. However, Basak and O'Connell had used well-learned verbal stimuli (digits), and the training was not adaptive or gamified to ensure engagement. The BirdWatch Game features qualities found to be effective in past cognitive training, including adaptive scaling difficulty (Boot et al., 2010; Payne et al., 2011; Brehmer et al., 2012; Cuenen et al., 2016) and computer-based gamification with novel stimuli that induce greater engagement and show transfer in older adults (Lampit et al., 2014; for meta-analyses, see Basak et al., 2020).

Considering that the BirdWatch Game is a working memory updating task, we hypothesize that the gray matter volumes of regions known to be related to working memory and cognitive control (e.g., frontal gyri, anterior cingulate cortex, premotor cortex, etc.) will positively predict its learning. The volumes of areas known to be related to learning in general (i.e., hippocampus and striatum) are likely to demonstrate a similar pattern. Additionally, considering the length of the training period utilized in this study, this study may reveal a differential relation between some of these examined volumes and early vs. late stages of learning. Specifically, the volume of the hippocampus may selectively relate to initial learning of the BirdWatch Game, considering its critical role in declarative learning (Burgess et al., 2002; Lim et al., 2020), and the theoretical contribution of episodic memory function to the cognition-dependent and strategy-dependent first and second stage of procedural learning (Ackerman, 1988; Beaunieux et al., 2006). Conversely, the volume of the striatum may selectively relate to later learning of the BirdWatch Game considering that region's contribution to procedural/automatized learning which occurs at later stages (Saint-Cyr and Taylor, 1992; Simonyan, 2019).

In terms of day-to-day predictors of task performance, contextual factors of sleep duration, stress, busyness, and physical and emotional wellbeing were examined as determinates of day-to-day performance on the BirdWatch Game learning. Sleep quality and duration are positively related to multiple cognitive abilities (Holanda Júnior and de Almondes, 2016; Lo et al., 2016; Rana et al., 2018; Zavecz et al., 2020), but stress negatively impacts working memory and cognitive control (Shields et al., 2016; Plieger and Reuter, 2020). Subjective wellbeing is also a positive correlate of working memory and cognitive control (Luerssen and Ayduk, 2017; Ihle et al., 2021). A secondary goal of this study was to examine how these contextual factors contribute to day-to-day performance on the training task. These measures, assessed at the onset of each training session, are hypothesized to predict overall performance during that session. Specifically, we hypothesize that stress and hours of sleep will have a strong aggregate effect if high stress or a few hours of sleep recur over several sessions, whereas wellbeing will relate positively to training performance. Additionally, considering past evidence (Festini et al., 2016), busyness may also relate positively to training performance.

Methods

Participants

A total of 55 older adults participated in a randomized clinical trial (RCT) contrasting different computerized cognitive training methodologies in healthy older adults (Basak,

NCT03988829), from which the present study drew data. Of the 43 participants randomized to the BirdWatch Game—Unity (BWGU) training, 37 participants ($M_{age} = 71.57$, $SD_{age} = 4.23$, 54% female) completed both baseline cognitive assessments and BWGU training period sufficient to be included in the present study. The remaining participants either explicitly ceased involvement in the study due to the outbreak of the COVID-19 pandemic in early 2020 or ceased responding to scheduling requests during the period of the pandemic.

Of the 37 participants included in this analysis, seven participants were unable to complete the structural MRI scans due to the periodic unavailability of MRI scanners due to the COVID-19 pandemic, as outlined above, resulting in a sample size of 30 participants ($M_{age} = 71.17$, $SD_{age} = 4.21$, 57% female) who contributed cognitive, MRI, and training data sufficient to be included in all the analyses presented below. Additionally, the difficulties of collecting data *via* in-person testing during the 2020–2021 COVID-19 pandemic resulted in a higher than expected number of participants with missing data ($n = 7$). Five participants were unable to contribute CRIq data due to technical difficulties arising from remote data collection during the period of the pandemic. Analyses presented in the following sections for which some participants were excluded due to missing data are explicitly noted.

Development of the BirdWatch game cognitive training program

At the core of this intervention program, titled the BirdWatch Game—Unity (BWGU), is the n-match paradigm, a modified n-back task in which participants must maintain and unpredictably update a number of items in their working memory simultaneously (Oberauer, 2006; Basak and Verhaeghen, 2011; Basak and O'Connell, 2016; O'Connell and Basak, 2018). In a typical n-back paradigm, participants are presented with a continuous sequence of individual stimuli and asked to compare the currently presented stimuli with the stimuli presented n items ago (Owen et al., 2005). Performing this task successfully requires participants to maintain the past *n* presented items within their working memory, continuously updating this information as new stimuli are presented (Jaeggi et al., 2010), and manipulating n in this paradigm thereby allows for the manipulation of participants' cognitive load.

The n-match paradigm (Basak and O'Connell, 2016) extends the traditional n-back paradigm by dynamically varying n during a single run of the task. This is accomplished by randomly presenting the stimuli in a set number of visuo-spatial contexts, and requiring participants to compare the currently displayed stimulus to the stimulus last displayed. For example, Basak and O'Connell (2016) utilized the numbers 1–9 presented in one to four different colors (the number of colors represented the n contexts of n-match task), and tasked participants with comparing the currently presented number with the most recent number presented in that same color. An earlier work by Basak and Verhaeghen (2011) utilized up to four different locations as contexts in an n-match paradigm to a similar effect. Due to the random presentation of context (color or location), participants are forced to actively maintain all *n* items within their working memory simultaneously and to unpredictably update this stored information, thereby increasing cognitive effort compared to a traditional n-back task where the n is fixed (Basak and Verhaeghen, 2011; Basak and O'Connell, 2016). The advantage of the n-match paradigm is that *n* can be dynamically varied by varying the sequence order of the context (e.g., Basak and O'Connell, 2016).

This intervention was based on the efficacy of executive function training in older adults of which working memory is an essential process (Basak et al., 2020), commonality of working memory issues as a subjective complaint in older adult populations (Newson and Kemps, 2006), and the theoretical efficacy of using working-memory-based training to address that complaint and contribute to general wellbeing (Luerssen and Ayduk, 2017). We elected to utilize the n-match training paradigm specifically as it has been shown to facilitate far transfer to measures of reasoning and episodic memory in older adults (Basak and O'Connell, 2016), and because the n-match tasks stressed working memory updating rather than just working memory span, which Miyake and Friedman (2012) identify as separate contributors to executive functioning.

The n-match paradigm described above was modified in several ways to produce the BWGU paradigm. First, to render the n-match paradigm more engaging, the paradigm was extensively gamified, i.e., modified to resemble a recreational video game. Simplified renderings of birds were used for individual stimuli, with trees in spatially distinct locations utilized as contexts (see Figure 1). Both bird stimuli and tree contexts are displayed on a rendering of an outdoor scene, selected to be both aesthetically pleasing and to reinforce the narrative that the BWGU training task is a "Bird Watching Game," as implied by the title of the task.

Additionally, we added game-like player feedback to BWGU in the form of a score display and a "reward" system. The score was calculated as follows:

$$\mathbf{Score} = \mathbf{100(Hit+CR)}\ \ \mathbf{50(Miss+FA)} + \mathbf{1000d'(7-MaxRT)}$$

In the above equation, *Hit* is the total number of hits from the previous block, *CR* is the total number of correct rejections from the previous block, *Miss* is the total number of misses from the previous block, *FA* is the total number of false alarms from the previous block, *d'* is the memory discriminability measure from the previous block, and *MaxRT* is the maximum allowed response time for the previous block (see below). While this scoring output is partially determined by performance metrics

FIGURE 1
A single trial from BWGU, depicting a four-context trial.

relevant to the goals of the present study, this score display was primarily implemented as an engagement tool that allowed participants to have a general sense of how their performance was progressing over time.

A "reward" system was implemented by the "unlocking" of new background images as participants met performance milestones, specifically whenever the performance threshold set by the program was increased (see below). This system was intended to somewhat reduce the monotony of performing the same task over multiple hours of training by periodically providing a different visual appearance over time, and to reinforce participant's success by tying this cosmetic change to performance milestones.

To further gamify this task, we implemented BWGU within the Unity game engine (Version 2018.4.2f1; 2018), a robust game development toolkit commonly used in independent game development. This allows BWGU to be deployed and run across multiple electronic platforms (i.e., Windows computers, Android and Apple phones, etc.) as if it were a recreational video game. As an added benefit, the Unity engine is sufficiently feature-rich and expandable to be comparable to data collection software more commonly used in cognitive science research (i.e., Eprime), which allowed for the collection of detailed performance metrics as described in the sections below.

Several methods of adjusting the difficulty of the BWGU task based on the participant's real-time performance were implemented within the paradigm based on past research, which implicates individualized-adaptive training methodologies as efficacious (Mihalca et al., 2011; Payne et al., 2011; Brehmer et al., 2012; Cuenen et al., 2016). First, BWGU continuously adjusts the number of contexts, *n*, utilized for a given block of trials based on participant performance in the previous block. Discrimination accuracy (*d'*) was utilized as the measure of participant performance and was calculated as $Z_{FA} - Z_{hit}$, where *FA* is the number of false alarms from the previous block, and *hit* is the number of correct identifications made in the last block. The 1/2N correction was applied to account for floor and ceiling effects (Macmillan and Creelman, 2005). The participant's *d'* for each block is compared to a performance threshold, d'_t, and *n* is incremented by 1 for the next block if *d'* is greater or equal. BWGU scales up to six contexts. Should a participant perform above threshold, the performance threshold is increased, and the number of contexts is reduced to one. This increase in d'_t is associated with the "reward system" with each increase in d'_t "unlocking" a new background display. The performance threshold begins at 0.6, and increments by +0.2 for each participant's success on an $n = 6$ block, to a maximum of $d'_t = 3$. This system allows the BWGU paradigm to scale up the difficulty in response to an individual participant's performance up to 72 times (six contexts by 12 increases in threshold) over the course of training (see Figure 2).

Additionally, the response time window in which a participant is able to enter a response to the current stimuli also scales in two ways with participant performance. By default,

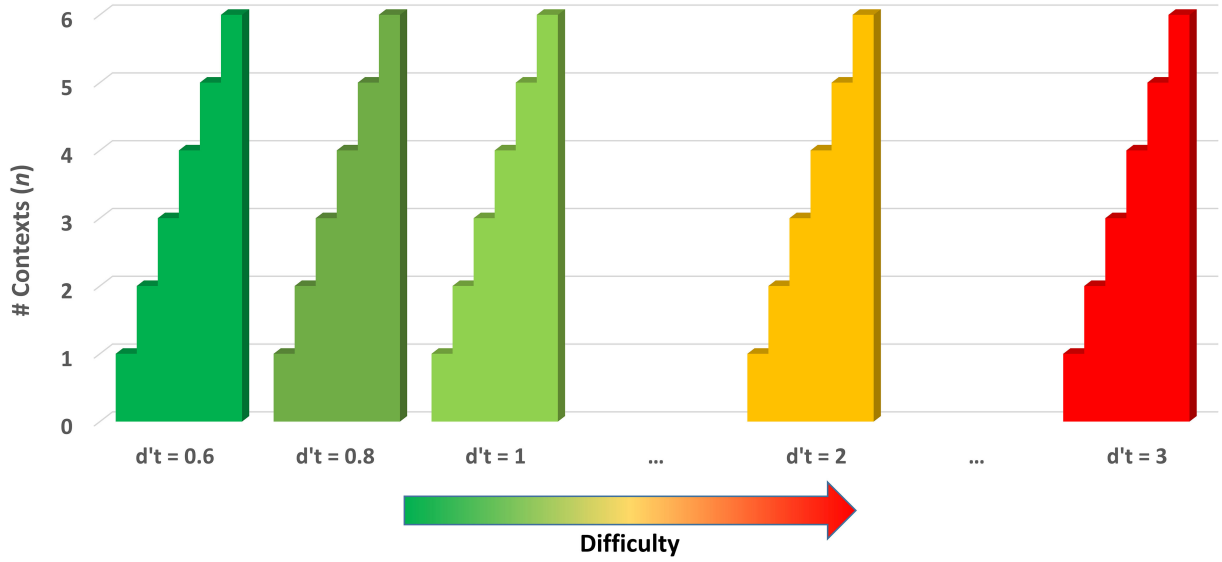


FIGURE 2
Depiction of overall difficulty progression by the number of contexts (*n*) and performance threshold (d'_t) in the BWGU paradigm.

participants have 5 s to respond to a new stimulus (i.e., *MaxRT* = 5 s). If an input is not detected in that time, that trial is marked as a "miss," and the task progresses to the next trial. For each 10% of the total expected training time elapsed, *MaxRT* is decremented by 0.5 s to a minimum of 1 s. Conversely, for every three consecutive failures to pass the performance threshold at the end of a block of trials, *MaxRT* is incremented by 0.5 s, to a maximum of 6 s. In this way, time pressure is both increased and decreased in line with the participant's performance and progress through training.

Implementation of BWGU in a multi-armed randomized controlled trial

The BWGU was utilized in two training arms of this RCT that contrasted various degrees of cognitive control over 20 h of training (Basak, NCT03988829). The two training arms of BWGU varied only in the sequence order of the context within a block, while all other features remained the same.

Recruitment

General inclusion criteria for the RCT were as follows: minimum age of 65 years, at least a 10th-grade education, learned English before the age of 5 years, and cognitively unimpaired (i.e., a Montreal Cognitive Assessment/MoCA score of 26 or greater; Nasreddine et al., 2005). Exclusion criteria included a history of cardiovascular disease other than treated hypertension, diabetes, psychiatric disorder, illness or trauma affecting the CNS, substance/alcohol abuse, and medication with anti-psychotics or hypnotics other than occasionally used at bedtime.

In addition to the above criteria, participants in the RCT were required to fulfill additional exclusion criteria in order to undergo the structural MRI portion of the study. Inclusion criteria for the MRI portion of the trial included right-handedness. Exclusion criteria for the MRI portion of the trial included metal medical implants, claustrophobia, and pregnancy. Initial recruitment for the RCT targeted only participants that fulfilled both the general and MRI inclusion/exclusion criteria outlined above. However, the onset of the COVID-19 pandemic in March of 2020 necessitated the expansion of the study to include participants who did not meet the criteria for MRI scans due to (a) high attrition of participants due to the pandemic, and (b) the necessity to conduct only remote cognitive testing between March 2020 and March 2021.

Training protocol and cognitive assessments at baseline

Participants in both BWGU arms were asked to train for 20 h over a period of 8 weeks on the BWGU paradigm. Participants were asked to train for 2.5 h each week, divided across two to three sessions. The training was performed at home using a 9.6' Android tablet computer provided to the participants, with the BWGU training program pre-installed on that device.

For the purpose of this longitudinal investigation, BWGU was configured to administer continuous blocks of 80 trials, with *n*, d'_t, and *MaxRT* modulated between blocks as described in

Section Development of the BirdWatch game cognitive training program. Between blocks, the BWGU training program pauses until the participant indicates they are ready to begin another block or chooses to exit the program. In the latter case, the current value of *n*, d'_t, and *MaxRT*, as well as the total training time completed, are saved by the program for use the next time the participant activates the training program. An additional feedback mechanism, a "progress bar," was added to the BWGU training program to aid participants in tracking their progress through training. This progress bar, which can be seen in the top center of Figure 1, fills relative to the participant's progression through the assigned 20 h of training, with the percentage of the bar filled reflecting the percentage of total training time elapsed.

Trial-wise performance data collected by the program includes participant accuracy, reaction time, and trial characteristics (switch trial and update trial). Block-wise performance data collected includes *Score*, *n*, d', and d'_t.

Cognitive reserve was assessed at baseline using the Cognitive Reserve Index Questionnaire (CRIq; Nucci et al., 2012). This self-report questionnaire assesses cognitive reserve as an aggregate effect of occupational, educational, and leisure activities over the lifetime, and has been demonstrated to both be independent of measures of general intelligence (Nucci et al., 2012) and reliable across a wide range of populations (Maiovis et al., 2016; Ozakbas et al., 2021).

Episodic memory measures administered at baseline and post-training included the Rey Auditory Verbal Learning Test (RAVLT; Bean, 2011) and the *Story Memory* sub-measure of the Mini-Mental State Examination (Folstein et al., 1975). The RAVLT is a word-list learning task of 15 that includes measures of simple learning, long-term memory (LTM) interference after distraction, LTM interference after delay, and multiple forms of LTM errors (source memory, semantic, and phonetic confusions). The Story Recall task is a modified word-list memory task in which the to-be-remembered items form a simple narrative separated into 34 distinct units. Participants are asked to read the story once, and then asked to recite it in as close to the original language as possible. An everyday test of memory was also administered, which included sub-measures of prospective memory, non-verbal recognition memory, and spatial-relational memory. However, the test proved infeasible to administer remotely, and as a result of this and the co-occurrence of the COVID-19 pandemic with data collection for this study, six participants were unable to contribute data for this everyday memory test. As a result of this, this test was dropped as an episodic memory measure in the analysis.

Reasoning measures administered at baseline and post-training included Visual Puzzles and Matrix Reasoning sub-measures of the Wechsler Adult Intelligence Scale, 4th edition (Drozdick et al., 2012). The Visual Puzzles test is a timed non-verbal reasoning test in which participants are presented with a series of puzzles of increasing difficulty. The Matrix Reasoning test is, similarly, a timed non-verbal reasoning test in which participants are presented with a series of incomplete visual patterns of increasing difficulty.

The current study used only the pre-training baseline assessments of the above-mentioned cognitive indices of far transfer (reasoning, episodic memory, and cognitive reserve).

MRI protocol

Baseline and post-training scanning protocols were conducted using a Siemens Magnetom Prisma scanner with a 32-channel head coil. High-resolution anatomical images were acquired using a transverse MPRAGE T1-weighted sequence with the following parameters: TR = 2,300 ms; TE = 2.26 ms; flip angle = 8°; acquisition matrix = 256 × 256; voxel size = 1 mm^3; 208 slices.

Specific information regarding the additional neuroimaging scans and behavioral assessments can be found in the preregistration for the RCT (Basak, NCT03988829). Data from these additional scans were not examined, as the current study specifically examined brain volume predictors of BWGU learning.

Daily survey of subjective wellbeing and sleep

To assess the impact of daily wellbeing on training performance-over-time, a short "daily survey" of subjective wellbeing and sleep measures was implemented in the BWGU training program. Participants were required to complete this survey each time they turn the program on, before their first block of training (see Figure 3).

The daily survey consists of a four-item Likert questionnaire on a 1–5 scale. Questions asked include (1) "How well did you feel in the past 24 h?" (2) "How stressed did you feel in the past 24 h?" (3) "How busy were you in the past 24 h?" and (4) "How was your mood in the last 24 h?" Questions 1 and 4 were presented on a scale from "1: very poor" to "5: very good," and questions 2 and 3 were presented on a scale from "1: not at all" to "5: very." Participant responses to questions 1 through 4 on this survey were taken as the Wellbeing, Stress, Busyness, and Mood variables, respectively. In addition to these Likert measures, participants were also asked to estimate their hours of sleep on the previous night, which was recorded as the Sleep variable.

Analysis

Calculation of learning rates

The *Difficulty Level* of each block was assessed by counting the number of times that the BWGU had adaptively increased the demands of the task based on the participant's performance prior to the beginning of that block (see Section Participants).

FIGURE 3
Screenshot of the Daily Survey Screen that appears just after the log-in screen in the BWGU.

This calculation can be formally represented as follows:

$$\textbf{Difficulty Level} = 6\frac{\mathbf{d_t'} - .6}{.2} + \mathbf{n}$$

In the above equation, d_t' represents the d-prime threshold of that block, and n represents the number of contexts for that block. Functionally, this results in the *Difficulty Level* for a block incrementing by +1 if either the number of contexts or the d' threshold has been updated since the previous training block. As the BWGU paradigm is designed to only adjust difficulty upward in response to player performance, we can correctly assume that any change in d_t' or n to reflect an increase in difficulty, and therefore the total number of adjustments equates to the total difficulty of the training block. Assigning the first block of training the *Difficulty Level* of 1 results in a range of 1–72 for this variable (see Figure 2).

In order to differentiate performance on training blocks of the same difficulty level, the *Difficulty Level* per block was multiplied by that block's unscaled accuracy (hits + correct rejections, range 0–80, chance performance = 40), to produce a *Simple Score* for each block. This *Simple Score* variable was used to calculate learning rates for each participant, as described below.

Past publications have used video game scores to calculate participant learning rates by fitting logarithmic curves to participants' scores over time, and taking the growth rate of that learning function as indicative of the rate of learning in older adults (Basak et al., 2011; Basak and O'Connell, 2016; Ray et al., 2017; Smith et al., 2020). Visual inspection of the *Simple Score* variable suggested that it followed a similar logarithmic pattern (see Figure 4), and so a similar method was employed in this study. The following logarithmic function was fit to each participant's *Simple Score* block-wise performance:

$$\mathbf{Y} = \mathbf{b0} + (\mathbf{b1} * \ln(\mathbf{t}))$$

In the above equation, t is the block of training (ordered sequentially, analog of training time/session), Y reflects the participant's *Simple Score* for a given t, $b0$ is the function's x-intercept, and $b1$ is the function's growth rate or slope. The growth rate of this function, as fitted to each participant's performance-over-time, was taken as that participant's *Overall Learning Rate*.

Compared to earlier studies that have used this method, the current study utilized a longer training intervention of 20 h (for an exception, see Basak et al., 2011, where training duration was 20 h). To account for this longer duration of the training, learning rates for early, middle, and late learning were calculated for each participant, corresponding to 1–5, 6–10, and 10–20 h of training, respectively, in addition to their *Overall Learning Rate*. The decision to define early, middle, and late learning in this way was based on a previous study where extensive practice on the n-back tasks in young adults stabilized after 5 h of training

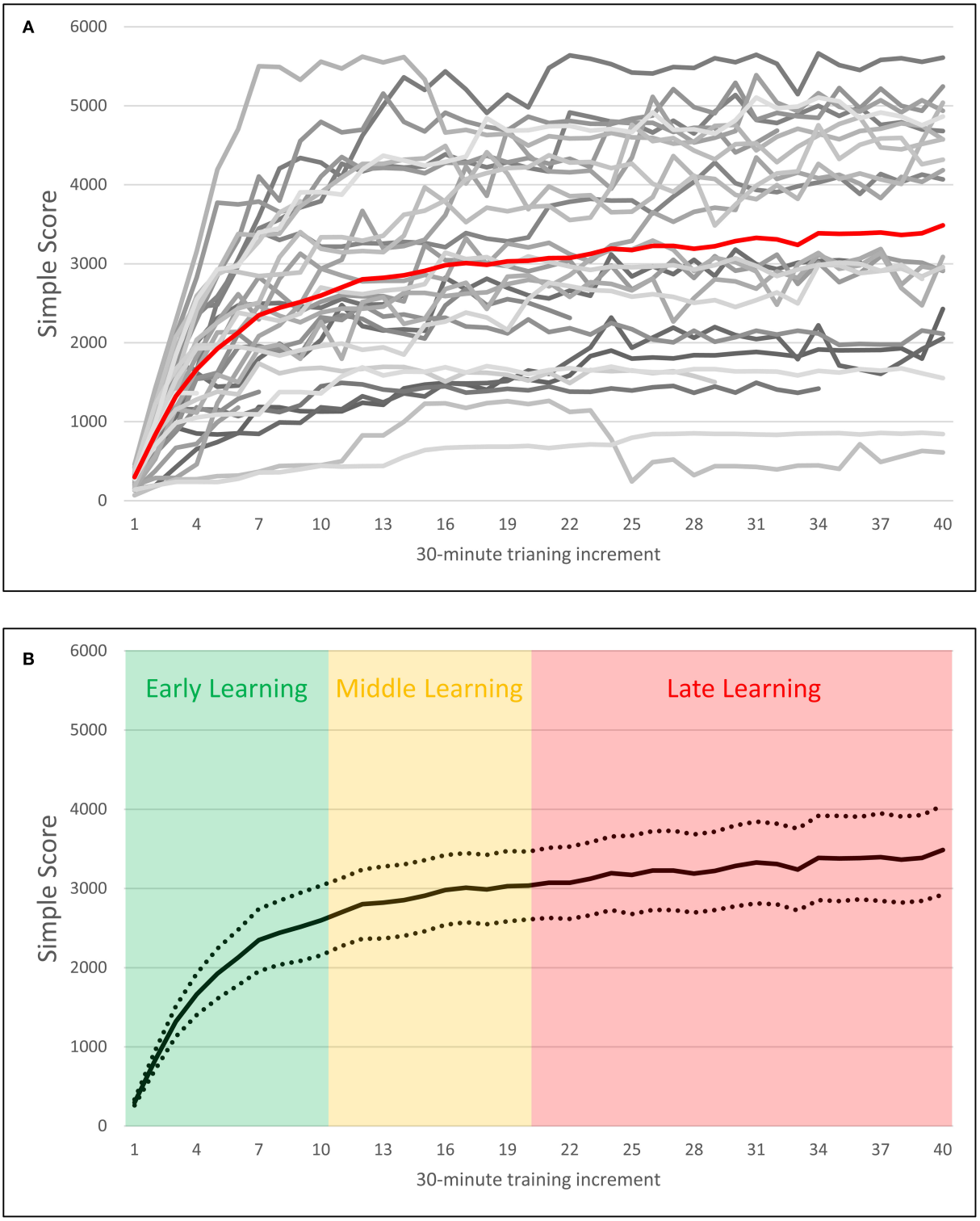


FIGURE 4
Plots of block-wise *Simple Score* by 30-min training increment over 20 h of training. **(A)** Depicts scores over time for individual participants represented in grayscale, with the average score over time plotted in red. **(B)** Depicts average scores over time with 95% confidence intervals, as well as demarcations of early, middle, and late training periods.

(Verhaeghen et al., 2004). As can be observed in Figure 4A, older participants also universally exhibited increasing performance across the first 5 h of training. Similarly, those participants who were able to reach asymptotic performance typically did so by the 10th h of training, as indicated by a relatively stable performance after 10 h of training. Based on these observations, hours 1–5 were designated as "early" training, and hours 10–20 as "late" training. The remaining period of hours 5–10 was designated as "middle" training, as the majority of participants appear to reach asymptotic performance within this range, but with specific achievement being time-variant.

An alternative approach to designating these training periods for the entire dataset would be to individually assign the early, middle, and late learning labels based on each individual subject's performance curve. However, we elected against this approach for two reasons. First, assigning learning periods across the whole group allows these data to be more readily comparable, a varying time period labeled as "early learning," for example, would make interpretation of results related to that training period problematic. Second, defining those periods for the entire dataset rather than per participant reduces the potential for unconscious coder bias during the coder's division of the learning period for each individual.

Based on visual inspection of the learning data (see Figure 4B), logarithmic functions were fitted to participants' early learning period, with linear functions fitted to their middle and late learning periods. As with the *Overall Learning Rate,* the growth rate of the log function fitted to early learning data was considered as each participant's *Early Learning Rate*. Similarly, the slope of the linear functions fitted to participants' middle and late learning data was considered as each participant's *Middle Learning Rate* and *Late Learning Rate*, respectively. Due to variance in total training time, only *Early Learning* could be fitted for all participants. *Middle Learning* could be fitted for 32 of the 37 participants, and *Late Learning* could be fitted for 31 of the 37 participants. Information regarding variance in training time and compliance can be found in Results Section BWGU adherence and training outcomes.

Calculation of cognitive measures

As mentioned above, episodic memory measures administered before BWGU training included the Rey Auditory Verbal Learning Test (RAVLT; Bean, 2011) and the Story Recall sub-measure of the Mini-Mental State Examination (Folstein et al., 1975).

The RAVLT includes multiple outcomes of episodic memory, where target list A is learned across five trials (A1–A5), followed by incidental learning of non-target List B (B1), followed by a surprise recall of target list A (A6) after the interference from the non-target list, and 30-min delayed memory recall (A7) and recognition test for the target list (recognition A). Recognition of the target list also included source monitoring errors on the recognition trial (recognition B), semantic errors in the recognition trial (recognition SA), phonetic errors in the recognition trial (recognition PA), and compound source-semantic and source-phonetic errors on the recognition trial (recognition SB and PB). To simplify the outputs, we calculated several aggregate measures from the RAVLT's raw output. Trials A1 through A5 were summed to produce a measure of overall learning (*Learning Total*). The difference between trial A5 and A6 was taken as a measure of interference cost (*Interference Cost*). The difference between trial A6 and A7 was taken as a measure of delay cost (*Delay Cost*). The sum of all errors on the recognition portion of the RAVLT (recognition B, SA, PA, SB, and PB) was summed into a single measure of recognition errors (*Recognition Errors*). These five aggregate measures, along with the total score on the *Story Recall* measure, constituted the episodic memory variables.

As mentioned above, reasoning measures administered before BWGU training included the *Visual Puzzles* and *Matrix Reasoning* sub-measures of the Wechsler Adult Intelligence Scale, 4th edition (Drozdick et al., 2012). Participants' total score on each of these respective measures constitutes the reasoning variables in this analysis.

Assessment of regional gray matter volumes

Cortical reconstruction and volumetric segmentation of the structural MRI images taken at baseline were conducted with the FreeSurfer 6.0 image analysis suite (Desikan et al., 2006; http://surfer.nmr.mgh.harvard.edu/). FreeSurfer 6.0 was selected over prior versions of FreeSurfer, as that version of the program has been demonstrated to significantly mitigate segmentation errors known to be present in previous versions (Brown et al., 2020; Srinivasan et al., 2020). To further lessen the impact of segmentation errors potentially resulting from Fressurfer's method of automated segmentation, aggregate volumes were used when appropriate, as described below.

Gray matter regions with established links to cognitive control, especially working memory updating and complex skill learning in older adults, were selected as regions of interest to reflect the cognitive demands of the BWGU; these regions included superior, middle, and inferior frontal gyri (Adólfsdóttir et al., 2014; Qin and Basak, 2020), middle temporal gyrus (Zhu et al., 2019), anterior cingulate cortex (Basak et al., 2011; Qin and Basak, 2020), and premotor cortex (Basak et al., 2011). Additionally, the volumes of the hippocampus and striatum were included, due to the known involvement of these regions' in declarative (Burgess et al., 2002; Lim et al., 2020) and procedural learning (Saint-Cyr and Taylor, 1992; Erickson et al., 2010; Doppler et al., 2019; Simonyan, 2019).

FreeSurfer volume outputs corresponding to each of these above regions were summed for each participant to produce an estimated volume of that region for that participant. Striatal volume (*Striatum*) was estimated by summing the

separate volume outputs for the caudate, putamen, and nucleus accumbens. Volume estimates of the inferior frontal gyrus (*IFG*) were created by summing the respective volume estimates for the pars opercularis, pars orbitalis, and pars triangularis. The rostral middle frontal and caudal middle frontal volumes estimates of the middle frontal gyrus (MFG) were summed into a single volume estimate of that region. Similarly, rostral anterior cingulate and caudal anterior cingulate volumes output by the program were summed into a single volume estimate of the anterior cingulate cortex (*ACC*). As FreeSurfer does not distinguish between premotor and supplementary motor volumes, the output volume of the precentral gyrus as a whole (*Precentral*) was utilized in this study. The FreeSurfer volume estimates of the superior frontal (*SFG*) and middle temporal gyri (*MTG*), as well as the *Hippocampus*, were used as outputs to represent those regions.

Results

BWGU adherence and training outcomes

All participants in the BWGU training arms were instructed to play 20 h of BWGU over the 2-month training period, but self-monitored and self-reported their training time for the duration of the intervention. As a result, a high amount of variance was observed in terms of total training time (M_{Time} = 17.35 h, SD_{Time} = 5.93 h). In total, 23 participants successfully reached 20 h of training time with the BWGU paradigm. Of those participants who did not complete the full 20 h of training, five participants explicitly discontinued training (M_{Time} = 3.48 h, SD_{Time} = 1.02 h). The remaining nine participants self-reported that they had completed 20 h of training time, but in fact had not when the electronic records of their training time were assessed (M_{Time} = 16.01 h, SD_{Time} = 2.95 h).

As stated above, participants were required to complete a daily survey of subjective wellbeing and sleep each time they activated the training program. On average, participants completed 29 surveys over the course of the training period (M_{Survey} = 29.22, SD_{Survey} = 14.11), with an average periodicity of one survey every 0.67 h of training ($SD_{SurveyTime}$ = 0.32). The number of surveys completed highly correlated with the total training time ($r = 0.67$, $p < 0.001$).

To assess if our variables of interest significantly differed between those participants who completed training and those who did not, we next ran a series of one-way ANOVAs comparing those participants who fully completed the training (*20+*), those who completed the training at under 20 total hours (*>20*), and those who discontinued training ("Discontinued"). Variables assessed in this way included age, *MoCA* score, years of education, *CRIq*, and all of our cognitive variables of interest (*RAVLT* sub-measures, *Matrix Reasoning*, *Visual Puzzles*, and *Story Memory*). These one-way ANOVAs demonstrated a marginally significant difference between the three completion groups in the *RAVLT Total Learning* and RAVLT *Interference Cost* measures: *Total Learning* $F_{(2/34)}$ = 2.83, p = 0.073; *Interference Cost* $F_{(2,34)}$ = 3.06, p = 0.06. *Post-hoc* comparisons using Tukey's method demonstrated that, in both cases, these effects were driven by marginal differences between the *discontinued* group and the other groups. The group that discontinued training demonstrated a marginally lower *Total Learning* than both the *20+* (p = 0.93) and *>20* (p = 0.76) groups, as well as a marginally higher *Interference Cost* than both the *20+* (p = 0.65) and *>20* (p = 0.76) groups. Those that completed training at greater or less than 20 total hours did not differ on these two measures (*Total Learning* p = 0.878; *Interference Cost* p = 0.958), and no other systematic differences in our variables of interest were detected between completion groups.

On average, participants reached level 51 of the BWGU paradigm, the coarsest measure of maximal attainment in this training paradigm, before ceasing training (M_{Level} = 51, SD_{Level} = 18.22), with subjects reaching maximal performance at ∼11 h of training on average ($M_{TimeHLR}$ = 11.36, $SD_{TimeHLR}$ = 5.86). A total of nine participants (24.3% of the sample) reached the maximum difficulty level allowed by the program (72) over the course of the training period. On average, participants completed ∼468 individual blocks of the BWGU paradigm throughout the training period (M_{Blocks} = 467.97, SD_{Blocks} = 262.23), with each block lasting an average of 2.25 min ($M_{BlockTime}$ = 2.25, $SD_{BlockTime}$ = 1.21). Predictably, both highest level reached and number of blocks completed highly correlated with the total training time: *HLR r(37)* = 0.5, p = 0.002; *Blocks r(37)* = 0.59, $p < 0.001$.There were no significant differences between the two BGWU arms regarding the total hours played [$t_{(36)}$ = 0.2, p = 0.84], highest level reached [$t_{(36)}$ = 0.79, p = 0.43], or number of blocks completed [$t_{(36)}$ = −0.29, p = 0.77]. A summary of participant training statistics can be found in Table 1.

Assessment of the relationship between cognitive reserve and cognitive ability prior to BWGU training

To assess if the cognitive reserve was related to baseline cognitive measures, we ran a series of partial correlations between the *CRIq* measure and the pre-training cognitive measurements (*RAVLT: Total Learning, Interference Cost, Delay Cost, Recognition Errors; Matrix Reasoning; Visual Puzzles; Story Memory*), controlling for *Age*. *CRIq* did not demonstrate any significant correlation with *RAVLT* sub-measures, *Total Learning* $r(29)$ = 0.16, p = 0.377; *Interference Cost* $r(29)$ = −0.15, p = 0.43; *Delay Cost* $r(29)$ = 0.21, p =0.267; *Recognition Errors* $r(29)$ = −0.32, p = 0.082, nor with *Matrix Reasoning*, $r(29)$ = −0.01, p = 0.964, or *Visual Puzzles*, $r(29)$ = 0.04,

TABLE 1 Summary statistics for demographic variables, cognitive measures, and the BirdWatch Game—Unity (BWGU) learning measures.

Measure	Mean (SD)
Demographics	
Age	71.57 (4.23)
Female	0.54
Education (years)	17.35 (3.15)
MoCA	27.89 (1.56)
Cognitive measures	
CRIq	130.66 (34.66)
RAVLT Learning Total	48.51 (12.25)
RAVLT Interference Cost	2.12 (1.95)
RAVLT Delay Cost	0.27 (1.54)
RAVLT Recognition Errors	2.03 (3)
Matrix Reasoning	16 (4.06)
Visual Puzzles	12.51 (3.88)
Story Memory	13.68 (5.52)
BWGU learning measures	
Time trained (hours)	17.35 (5.93)
Blocks completed	467.5 (262.23)
HLR	51 (18.2)
Overall learning (growth)	639.42 (348.27)
Early learning (growth)	712.67 (401.74)
Middle learning (slope)	3.08 (4.51)
Late learning (slope)	0.71 (2.02)

$p = 0.827$, or *Story Memory*, $r(29) = 0.02$, $p = 0.93$. These results indicate that cognitive reserve, as measured by the CRIq, is unrelated to pre-training (baseline) cognitive ability in this study.

Effect of individual differences in baseline cognition and cognitive reserve on BWGU learning

To assess the impact of variance in baseline cognitive measures on learning of the BWGU task, a series of stepwise multiple regressions were conducted with participants' learning variables (*Overall, Early, Middle,* and *Late Learning*) as dependent variables. In each of these regressions, the cognitive predictors (*RAVLT*: *Learning Total*, *Interference Cost*, *Delay Cost*, and *Recognition Errors*; *Story Memory*; *Matrix Reasoning*; and *Visual Puzzles*) were entered in a stepwise fashion until only significant predictors remained.

Overall Learning was found to be marginally predicted by a model containing only *Story Memory*, $R^2 = 0.14$, $F_{(1,35)} = 5.86$, $p = 0.021$, *Story Memory* $\beta = 22.34$, $t_{(35)} = 2.21$, $p = 0.034$. Similarly, *Early Learning* was found to be significantly predicted by a model containing only *Story Memory*, $R^2 = 0.16$, $F_{(1,35)} = 6.48$, $p = 0.015$, *Story Memory* $\beta = 28.78$, $t_{(35)} = 2.55$, $p = 0.015$. Models were not successfully fitted to *Middle* or *Late Learning*, as no combination of the examined predictors produced a model with $p < 0.1$. To assess if the above relationships co-varied with *Age*, we conducted a series of two-step hierarchical regressions predicting *Overall Learning* and *Early Learning*, respectively. *Age* was entered as a covariate in step 1 of these analyses, with *Story Memory* entered in step 2. In the analysis correcting for age, *Overall Learning* was found to be marginally significantly predicted by a model containing both *Age* and *Story Memory*, $R^2 = 0.16$, $F_{(2,34)} = 3.26$, $p = 0.051$. Within the model, only *Story Memory* was significant, $\beta = 22.34$, $t_{(34)} = 2.21$, $p = 0.034$. Similarly, *Early Learning* was found to be significantly predicted by a model containing *Age* and *Story Memory*, $R^2 = 0.31$, $F_{(2,34)} = 7.72$, $p = 0.002$, where both *Age* and *Story Memory* significantly contributed to that model in the expected directions, *Age:* $\beta = -38.23$, $t_{(34)} = -2.78$, $p = 0.009$; *Story Memory:* $\beta = 22.37$, $t_{(34)} = 2.22$, $p = 0.033$.

Next, a series of regressions were used to assess the influence of cognitive reserve (*CRIq*) on BWGU learning. As with the assessment of cognitive predictors, one regression was performed with *Overall, Early, Middle,* and *Late Learning* as respective dependent variables. In these regressions, *Age* was entered in step 1 as a control variable, followed by *CRIq* in step 2 as the variable of interest. *CRIq* did not significantly predict *Overall Learning*, $R^2 = 0.06$, $F_{(1,30)} = 1.74$, $p = 0.197$, or any of the discrete learning periods examined [*Early Learning:* $R^2 = 0.03$, $F_{(1,25)} = 1.04$, $p = 0.316$; *Middle Learning:* $R^2 = 0.01$, $F_{(1,35)} = 0.16$, $p = 0.69$; *Late Learning:* $R^2 = 0.01$, $F_{(1,25)} = 0.23$, $p = 0.639$]. Note that the combination of between-subject variance in training time and the lack of CRIq data for some participants resulted in these analyses having substantially lower n as compared to the analysis of cognitive predictors (overall and early learning: $n = 37$ for cognitive predictors, $n = 32$ for CRIq; middle learning: $n = 32$ for cognitive predictors, $n = 27$ for CRIq; late learning $n = 31$ for cognitive predictors, $n = 27$ for CRIq).

Effect of individual differences in gray matter volume on BWGU learning

To assess the impact of variance in regional gray matter volumes on learning of the BWGU task, a series of multiple repressions were conducted with participants' learning variables (*Overall, Early, Middle,* and *Late Learning*) as dependent variables. In each of these regressions, the gray matter volumes from the baseline imaging session (left and right *SFG*, *MFG*, *IFG*, *ACC*, *Precentral*, *MTG*, *Hippocampus*, and *Striatum* volume) were entered in a stepwise fashion until only significant predictors remained. These analyses produced a significant

model of *Early Learning* ($R^2 = 0.16$, $F_{(1,28)} = 5.36$, $p = 0.028$), where the volume of the left *IFG* was the sole contributor ($\beta = 0.109$, $t_{(28)} = 2.31$, $p = 0.028$). To evaluate if the relationship between left *IFG* volume and *Early Learning* is significant even after controlling for nuisance variables, a stepwise regression was conducted with *Early Learning* as the dependent variable, Age and estimated total intracranial volume (eTIV) as covariates in step 1, and the left IFG volume in step 2. This resulted in a significant model that predicted *Early Learning* ($R^2 = 0.28$, $F_{(3,26)} = 3.42$, $p = 0.032$), with both *Age* and left *IFG* volume as marginally significant predictors, *Age*: $\beta = -32.75$, $t_{(27)} = -1.96$, $p = 0.061$; left *IFG*: $\beta = 0.1$, $t_{(27)} = 1.9$, $p = 0.064$.

Combined effects of cognitive and gray matter volume predictors on BWGU learning

The above analyses identified one significant cognitive predictor (*Story Memory*) and one significant brain structure predictor (the volume of the left IFG) of early learning. The influence of *Story Memory* on Early Learning contributed to its influence on Overall Learning. To evaluate the combined effects of these predictors, we conducted two stepwise regressions with *Overall* and *Early Learning* as respective dependent variables. In both regressions, *Age* and *eTIV* were entered in step 1 as control variables, and both *left IFG* and *Story Memory* were entered as variables of interest in step 2. This combinatorial model was found to significantly predict *Early Learning*, $R^2 = 0.44$, $F_{(4,25)} = 4.82$, $p = 0.005$, with both *Story Memory* and *left IFG* contributing significantly to the model, *Story Memory*, $\beta = 35.59$, $t_{(25)} = 2.6$, $p = 0.016$; *left IFG*, $\beta = 0.11$, $t_{(25)} = 2.25$, $p = 0.033$. This combined model was also found to significantly predict *Overall Learning*, $R^2 = 0.38$, $F_{(4,25)} = 3.86$, $p = 0.014$, with both *Story Memory* and *left IFG* contributing significantly to the model, *Story Memory*, $\beta = 41.73$, $t_{(25)} = 3.44$, $p = 0.002$; *left IFG*, $\beta = 0.09$, $t_{(25)} = 2.02$, $p = 0.054$.

Impact of daily context on daily BWGU performance: A time series forecasting analysis

To assess the individual-level influence of daily psychosocial factors on performance-over-time, we ran a series of auto-regressive integrated moving average (ARIMA) analyses using Simple Score as the dependent variable, Training Day as the indexing variable, and Wellness, Stress, Busyness, Mood, and Sleep as independent variables. This analysis was run independently for each participant, allowing for individual assessment of the impact of each moderator on performance-over-time. A total of three participants entered the same response for one or more of the psychosocial context questions for the entire duration of their training, resulting in those psychosocial variables exhibiting zero variance for those participants. Thus, these invariant variables were removed from those participants' models.

These ARIMA analyses were accomplished using the "forecast" package (Hyndman and Khandakar, 2008; Hyndman et al., 2021) for R (R Core Team, 2013). Instead of setting the AR, I, and MA, parameters of the ARIMA models *a piori*, the auto.arima function of the "forecast" package was used to procedurally select the ARIMA model that best fitted each participant's time series. This auto-ARIMA approach examines all possible ARIMA models within the bounds specified, and selects a final model based on the Akaike Information Criterion (AIC), which is a model criterion that accounts for both goodness-of-fit and parsimony of the model (Akaike, 1973, 1987; Sawa, 1978; Bozdogan, 1987, 2000). Maximum parameter bounds for these auto-ARIMA analyses were set to $AR \leq 5$, $I \leq 1$, $MA \leq 5$.

Individual ARIMA models of best fit: Prior performance forecasting future performance

The ARIMA models were successfully fit for 34 participants. The ARIMA models did not fit the remaining three participants due to a conjunction of low training time (all three participants discontinued the study prior to completing 5 h of training) and a sparsity of daily survey responses (i.e., longer play sessions resulting in fewer survey prompts occurring during training).

High heterogeneity was observed in the models of best fit across these 34 participants. Ten distinct models were found to be the model-of-best-fit for at least one participant. Of these 10 models, the most common models of best fit were the $AR = 0$, $I = 0$, and $MA = 0$ model ("000") and the $AR = 0$, $I = 1$, and $MA = 0$ model ("010"), each fitting $n = 7$ participants and together fitting 14 (41%) participants. Both models-of-best-fit feature *AR* and *MA* terms of 0, indicating that the performance of 14 (out of 34) participants on a given day was not strongly influenced by either their prior performances or the moving average of error of their performance on previous days. A summary of all models found to fit at least one participant can be found in Table 2.

For the remaining participants, 17 (50%) participants' data were best fitted by a model with an *AR* term of one or higher ($M_{AR} = 1.41$, range 1–3, see Figure 5), indicating a predictive influence of previous days' performance on the current day's performance. Five participants (14.71%) were fitted by a model counting an *MA* term of 1, indicating that, for those participants, current performance on the BWGU task was predicted by the error term of their previous day's

performance. Eighteen participants' data (52.94%) were fit by a model that included an integration (*I*) term of 1, indicating that these participants' performance-over-time exhibited non-stationarity which first-order integration was able to account for (Papoulis, 2002). In total, the performance-over-time of 19 participants (55.88%) was predicted by their previous day's performance, as indicated by a model-of-best-fit which included a non-zero *AR* and/or *MA* term.

TABLE 2 ARIMA models found to significantly explain performance-over-time in at least one participant, grouped by number of occurrences.

Model	*AR* term	*I* term	*MA* term	*n*
"000"	0	0	0	7
"100"	1	0	0	5
"200"	2	0	0	2
"300"	3	0	0	1
"010"	0	1	0	7
"110"	1	1	0	4
"210"	2	1	0	2
"011"	0	1	1	2
"111"	1	1	1	2
"211"	2	1	1	1

Individual ARIMA models of best fit: Wellbeing and sleep as predictors of BWGU performance-over-time

As with the model terms of each participant's model-of-best-fit, the value and significance of the psychosocial context moderators and sleep on each participant's performance-over-time also demonstrated notable heterogeneity. Stress significantly predicted performance-over-time at $p < 0.05$, in seven participants (21.21% of the sample), and was found to be the most common single contextual predictor. Wellness significantly predicted performance-over-time on the BWGU task at $p < 0.05$ in four participants (12.12% of the sample). Busyness significantly predicted performance-over-time at $p < 0.05$ in five participants (12.5% of the sample). Mood significantly predicted performance-over-time at $p < 0.05$ in four participants (12.12% of the sample). Sleep also significantly predicted performance-over-time ($p < 0.05$) in three participants (8.82% of sample).

In total, 17 (50%) of the sample demonstrated performance-over-time which was predicted by one or more of the examined psychosocial context variables and sleep, whereas the remaining 17 (50%) participants demonstrated no such relationship.

For participants who exhibited significant relationships between the psychosocial context variables (including sleep) and BWGU performance-over-time, all five variables demonstrated a negative relationship with performance: *Wellness* $M_\beta =$

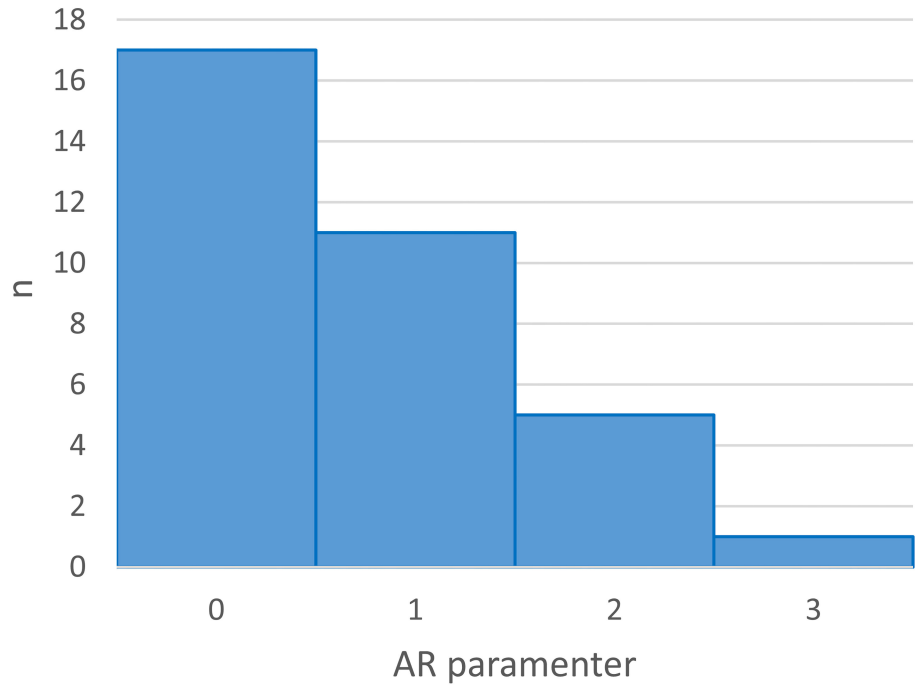


FIGURE 5
Histogram of AR term values in individual participant's ARIMA model-of-best-fit.

-200.67, $\sigma_\beta = 192.43$; *Stress* $M_\beta = -1{,}012.94$, $\sigma_\beta = 1{,}175.51$; *Busyness* $M_\beta = -201.64$, $\sigma\beta = 295.99$; *Mood* $M_\beta = -292.19$, $\sigma_\beta = 365.52$; *Sleep* $M_\beta = -775.64$, $\sigma_\beta = 1{,}063.64$. These results demonstrate a highly individualized effect of the examined psychosocial variables on BWGU performance-over-time, including half of our sample for whom performance does not appear to be influenced by the psychosocial context variables examined. Full model reports for each participant can be found in the Supplementary material.

Discussion

The present study was designed to investigate the cognitive and brain structure correlates of learning a novel gamified computerized working memory task (the BWGU), in order to determine if this game is used as an intervention, what is its potential for far transfer to reasoning and episodic memory and to induce brain plasticity. The results presented above identify one cognitive measure and one structural volume predictor of learning on the BWGU, even after controlling for individual differences in age, specifically of learning within the first 5 hours of the task. In terms of cognitive performance, participant's score on *Story Memory*, a measure of episodic memory, positively related to the participants' learning rates during the first 5 h of practice on the BWGU. In terms of structural volume predictors, estimated gray matter volumes (GMVs) of the participants' left inferior frontal gyrus (IFG) were predictive of learning of the BWGU task during the same period, even after controlling for age.

The *Story Memory* task is a modification of a word-list-recall episodic memory task, with the world list forming a narrative of a coherent episode (Folstein et al., 1975). The strong narrative aspect of the *Story Memory* paradigm may partially explain why performance on that measure was predictive of performance on the BWGU task specifically. One of the modifications made to the BWGU paradigm to increase its efficacy over a traditional n-match was the application of the "bird watching" narrative to the task. It is possible that this narrative operated as a contextual framing device that facilitated performance on the task. If that is the case, the ability to represent and elaborate on this narrative in a way that supports memory, indicated by higher *Story Memory* performance, may have allowed participants to learn the BWGU task at an increased rate.

A past study by Beaunieux et al. (2009) found a somewhat similar pattern of results regarding episodic memory and novel task learning to what was found in the present study. Beaunieux et al. (2009) found that measures of both working and episodic memory predicted a successful acquisition of a novel reasoning task (the Tower of Toronto, Saint-Cyr et al., 1988) over four training sessions. Additionally, Beaunieux et al. (2009) found that episodic memory deficits in older adults (aged 65+ years) in particular, as compared to their younger adult cohort, were negatively predictive of learning on the reasoning task. From this perspective, the results of this study can be interpreted as a specific case of cognitive reserve: degree of retention (i.e., degree of reserve) of episodic memory function, measured in this study by the *Story Memory* measure, may have facilitated learning of the BWGU Task, much as Beaunieux et al. (2009) theorized that preserved episodic memory function did on the Tower of Toronto task in their study.

If the relationship between *Story Memory* and BWGU learning allows for speculation for transfer from training, it is possible that training older adults on BWGU, especially for 5 h or so, may engender transfer to *Story Memory*. This hypothesis is supported by Basak and O'Connell (2016), where 5 h of unpredictable n-match training engendered greater transfer to *Story Memory* recall than the predictable n-match training in older adults. Importantly, faster learning rates were related to greater improvements in *Story Memory*.

Regarding gray matter volume, left inferior frontal gyrus volume significantly predicted learning of the BWGU. As with the *Story Memory* task, left IFG volume was not only found to specifically predict learning during the early phase of the training (hours 1–5), but also significantly contributed to a model predictive of overall learning along with the *Story Memory* measure. Considering the IFG's well-documented role in language processing (Hagoort, 2013; Fedorenko and Thompson-Schill, 2014), the conjunction of left IFG volume and *Story Memory* performance in predicting BWGU task learning strongly suggests that language processing contributes to the learning of the BWGU task. This is a plausible relationship if it is assumed that participants tended to use a verbalization or narrative-based strategy to aid in learning the BWGU task, such as assigning names to the otherwise un-named bird stimuli or applying/embellishing a narrative as a framing device to aid in memory and retrieval of the most recent bird stimuli observed. However, as no strategy self-reports were collected from participants for this study, we cannot assume this is the case. In the absence of confirmation of a language-based strategy for engaging with the BWGU task, exactly how individual differences in language processing would contribute to the learning of the BWGU task remain nebulous.

The inferior frontal gyrus is not, however, exclusively dedicated to language processing: there is ample evidence that it is involved in expressing cognitive control over memory processes more generally. A recent fMRI study by Qin and Basak (2020) found that the IFG is activated not only in younger but also in middle-aged and older adults during the unpredictable two-match task, where digits needed to be retrieved and continuously updated, along with other frontal and parietal regions that are implicated in cognitive control and working memory. Badre and Wagner (2007) concluded based on a review of the literature available at the time that IFG is specifically involved in enforcing cognitive control on the memory retrieval process, a capability essential to the expression

of language but not unique to that process (Fedorenko and Thompson-Schill, 2014). A model proposed by Hagoort (2013) specifies that the IFG serves to integrate information from regions of the brain involved in attentional, integrational, and memory processes in a way that allows for precise control of language. This body of work suggests that the IFG is heavily involved in the cognitive control processes of memory retrieval and updating, which are generalizable to language and other tasks. From this perspective, the observed relationship between greater gray matter volume in the left IFG and faster learning of the BWGU task can be interpreted not as dependent on language processing specifically, but that individuals with greater left IFG volume exhibit better cognitive control over their memory retrieval processes during training, thereby producing swifter learning of the task.

Importantly, even when considered together, these predictors (*Story Memory* and Left IFG) were independent contributors to the learning rate across the 20-h training session, even after accounting for age. They were also predictive of learning rates within the first 5 h of training. No significant predictors of the "middle" (hours 5–10) or "late" (hours 10–20) period of training were identified. Model fit and significance were greater when fitting the *Story Memory* + *IFG* model to *Early Learning* compared to *Overall Learning* ($\Delta R^2 = 0.06$), which suggests the pattern seen in overall learning may in fact be driven by the contribution of early learning to that variable. Indeed, a simple linear regression confirms that variation in *Early Learning* significantly explains ~68% of the variance in *Overall Learning* ($R_2 = 0.68$, $p < 0.001$), with another 17% of the variance being accounted for when *Middle* and *Late Learning* periods are added to the model ($R_2 = 0.85$, $p < 0.001$). These results would appear to confirm that learning within the first 5 h of training on the BWGU was the primary determinant of overall learning on that task.

The above relationship confirmed, why then were the observed structural and cognitive predictors of learning of the BWGU not related to learning rates in hours 5–10 or 10–20 of the training? The learning model proposed by Ackerman (1988) states that the first phase of learning is primarily determined by cognitive factors, with later learning primarily determined by the development of strategy and automatization of task-relevant responses. Considering that the potential predictors of learning that were examined in this study consisted of (a) cognitive predictors and (b) gray matter volume of regions related to the training task and cognitive predictors, it is no surprise then that any relationship uncovered would pertain to the early learning period specifically. The present study did not assess strategy formation or use by participants, and as such does not include a variable with sensitivity to Ackerman's strategy-dependent second phase of learning. The automatization-dependent third stage of Ackerman's model predicts stability of performance but improvement of response time on time-sensitive tasks. This flattening of performance is likely captured in the "late" learning period of the present study, defined by asymptotic performance on the BWGU tasks, but again no time-based variables sensitive to the development of automatized processing were examined in the analysis presented here. In short, strategy-based learning and automatization may well have been facilitated over 20 h of training on the BWGU task, but the game score analyzed here was not sensitive to those processes. This is not to say that this study's findings related to early learning are spurious. Rather, it should be recognized that variance in individual learning rates from strategy-based or automatic processes, both of which hypothetically contribute to later learning, are likely not accounted for in these analyses due to predictor and outcome variables utilized in this study.

Ackerman (1988) model of procedural learning offers an explanation as to why cognitive predictors of early learning were found in this study generally, but not why episodic memory measure and left IFG volume specifically predicted early learning of the BWGU task. Taken together, these predictors appear to reflect participants' ability to apply cognitive control to memory retrieval and, as needed, update the memory to encode it even for information that is tracked over a few seconds. As discussed above, aspects of the BWGU task itself, such as heavy emphasis on working memory updating, incorporating narrative framing device, as well as the known sensitivity of the *Story Memory* measure to n-back-based training (Basak and O'Connell, 2016) may well account for this. However, past work by Beaunieux et al. (2006) identified both episodic memory and cognitive control as indicative of learning a reasoning task (the Tower of Toronto). Beaunieux et al. (2006) concluded that episodic memory and executive function contributed to the first stage of learning in Ackerman's model. While the authors do not fully support that position based on the evidence provided by Beaunieux et al. (2006) that a similar pattern of predictors was found to relate to early learning on both the Tower of Toronto and the BWGU task suggests that these results might be generalizable beyond these select tasks, which is certainly worthy of future study. This study showed that in older adults who trained on a novel gamified, individualized-adaptive working memory updating intervention, the BirdWatch Game—Unity, for about 20 h, individual differences in a measure of episodic memory and the volume of left inferior frontal gyrus predicted individual's learning rate. These relationships were specifically applicable to the early phase of novel game learning, where individuals display rapid gains in game performance.

Importantly, neuro-cognitive predictors of skill learning on a task, such as BWGU, can inform us about the potential transfer mechanisms that can result from training on such skills. Another significance of this study is the potential identification of individuals who may benefit most from BWGU training.

Notably, the included measure of cognitive reserve (CRIq) did not reliably predict overall learning of the BWGU task, nor learning in any of the discrete training periods examined. This is perhaps not surprising as the cognitive reserve is typically

conceptualized as a protective factor (Tucker and Stern, 2011; Opdebeeck et al., 2016), rather than a factor that facilitates cognitive function, and the existing evidence linking cognitive reserve to task learning is somewhat weak (Lojo-Seoane et al., 2020). This is not to say that the study has definitively produced no evidence of reserve contribution to the learning of the BWGU task: as mentioned earlier, the observed relationship between *Story Memory* recall and BWGU learning may well be evidence of cognitive reserve, especially considering the degree of decline in episodic memory typically observed in older adults (Park et al., 2002; Rozas et al., 2008). A similar argument can be made regarding brain reserve. However, in the absence of cognitive or brain structure measurements taken from these participants earlier in life, these reserve arguments cannot be directly supported. Importantly, cognitive reserve is typically indexed by measures of life-time cognitive activity and educational attainment, and has been found to interact with cognitive training-related gains in cognition in healthy aging (for a meta-analysis, see Basak et al., 2020). It can be concluded, however, that cognitive reserve as measured by the CRIq as a sum of educational attainments and self-report aggregate of life experience does not relate to learning of the BWGU task.

The second goal of this study was to determine whether fluctuating psychosocial context variables and sleep duration influenced performance-over-time on the BWGU task. The most general hypothesis that sleep and the psychosocial variables examined would influence performance-over-time was demonstrably true for 50% of the sample, or 17 total participants, while the other 50% of the sample demonstrated no such relation. This, obviously, limits the conclusions we can draw based on this evidence. We cannot declare that a random participant from this sample would be more likely than not to be affected by one or more of the examined psychosocial context variables, due to simple probability. However, this result still allows for some definite conclusions to be drawn.

First, that performance-over-time of 50% of the sample of this study was influenced by at least one of the daily survey measures (that is, sleep, stress, busyness, mood, or wellbeing) is far from a negligible fraction. Indeed, if we assume that these results are generalizable, then it is fairly likely that performance-over time on the BWGU task will be influenced by one or more of these factors for a given participant. Additionally, there are likely undetected moderators which partially determine whether a given participant's performance is influenced by a given psychosocial context variable or sleep, which are important to further investigate considering how pervasive the influence of these psychosocial context variables and sleep are on cognition. Considering the well-documented negative impact of disrupted sleep (Holanda Júnior and de Almondes, 2016; Lo et al., 2016; Rana et al., 2018; Zavecz et al., 2020) and high stress (Shields et al., 2016; Plieger and Reuter, 2020) on cognitively demanding tasks in the real world, understanding what variables may moderate this relationship is of substantial real-world importance. The results of the present study indicate that sleep and the psychosocial context variables examined in this study can have an impact on the performance and learning of complex tasks, which is warranted enough for further investigation.

Second, while the generalization of these results is problematic, these models do offer significant explanatory power with regard to each individual participant. This has potential utility within the cognitive training domain as a method of assessing the individual needs of a participant, and providing cognitive training that is individually adaptive to those needs. Accurate models were fitted for participants who completed as little as 3 h of training, and for all participants who completed more than 5 h of training. Within the timescale of a long-term cognitive intervention, which typically involves 10-20 h of training (Basak et al., 2020), an analysis like the one performed in this study could be conducted with sufficient remaining time to provide individuated participant feedback or adjust the prescribed training, to account for any significant psychosocial effects observed. This is an alternate approach to individualized-adaptive training to the closed-loop strategy implemented in the design of the BWGU paradigm, where training difficultly was manipulated with respect to performance metrics (block-wise d' and consecutive failures), but not daily sleep or perceived wellbeing. Our current approach is agnostic to idiosyncratic influences on individual subjects, under the assumption that such sporadic daily influences are reflected in each participant's overall performance. Identifying and accounting for specific factors influencing performance-over-time, which the method of analysis presented in this study could facilitate, may serve as an effective additional method of adaptive training independent of the performance-focused method implemented in BWGU. Importantly, findings from the time series forecasting analysis provide evidence for why the individualized-adaptive approach to training has been generally successful at inducing positive training outcomes (Payne et al., 2011; Brehmer et al., 2012; Cuenen et al., 2016). A wide array of patterns of psychosocial influence were observed even within the age and geographically restricted sample utilized in this study, and it can be assumed with some confidence that individuals undergoing any form of cognitive training or intervention are subject to a similarly wide array of moderating influences.

The analysis presented in this study also demonstrated that 50% of the sample ($n = 17$) exhibited performance-over-time that was reliably predictable by previous performance, either through direct auto-regression of past performance onto a given day's BWGU score, or *via* a moving average of error terms. The finding that for 50% of our sample, current BGWU performance was not reliably predictable from past performance is interesting, as it suggests that other factors are primarily responsible for performance-over-time in this large proportion of the sample. As discussed, psychosocial context variables demonstrably accounted for variance in performance in half of our sample, which includes 11 of the 17 participants for

whom past performance did not relate to performance-over-time. However, this still leaves six participants for whom none of the examined variables, including their own performance, was predictive of variability in performance-over-time. The only conclusions that can be drawn about what these other factors might be are that they (a) have periodicity longer than the training period observed or (b) are transient events, as otherwise evidence of any such predictable influence would be detectable in the auto-regressive or moving average analysis. In light of these findings, it is clear that individual influences on performance-over-time on a complex task like the BWGU task are highly varied, and that they can be very influential. Further investigation of how these individual-level factors can be identified, modeled, and accounted for can only be a boon to efforts to develop efficacious, individualized cognitive interventions.

As already mentioned, the design of this study limits some of the conclusions we are able to draw from these results, and these design limitations can be improved upon in future iterations of this work. First, the present study did not take participant strategy into account. This is a particularly pertinent limitation to the findings of this study considering (a) the possibility that participants were utilizing a verbalization or narrative-based strategy to aid learning of the BWGU task, and (b) the theoretical relevance of strategy generation toward procedural task learning. A *post-hoc* self-report could potentially allow for insight into the effect of strategy on BWGU learning; however, this self-reported approach would need a much larger sample size of 250 or more given the variability of self-generated strategy reports and associated variables of interest, such as personality factors (e.g., openness to experience), cognitive flexibility, IQ, etc. Such a research agenda is challenging to implement in cognitive interventions that last for months and include brain measures. Another approach to studying the role of strategy could be a strategy manipulation applied *via* varied participant instructions, although this would require an in-lab intervention and a much larger multi-arm RCT that would have similar limitations of the feasibility of study implementation in terms of time and resource as described before. Second, the design of the present study did not allow for a detailed investigation of the influence of cognitive/brain reserve on learning of this task, beyond the retroactive self-report measure utilized by the CRIq. Addressing this shortcoming is somewhat difficult: A longitudinal approach by which trajectories of cognitive/neurological change over time could be calculated before the training period began could potentially enlighten and specify the reserve-learning relationship, but this would require a major expenditure of time and resources to accomplish.

Conclusion

This study showed that in older adults who trained on a novel gamified, individualized-adaptive working memory updating intervention, the BirdWatch Game—Unity, for about 20 h, individual differences in a measure of episodic memory and the volume of left inferior frontal gyrus predicted individual's learning rate. These relationships were specifically applicable to the early phase of novel game learning, where individuals display rapid gains in game performance. These predictors appear to reflect participants' ability to apply cognitive control to episodic memory functions, especially memory retrieval and subsequently updating the memory to encode it even for information that is tracked over a few seconds as in BWGU. Importantly, neuro-cognitive predictors of skill learning on a task, such as BWGU, can inform us about the potential transfer mechanisms that can result from training on such skills. In fact, prior research in older adults has shown that just 5 h of training on working memory updating, where stimulus sequence appeared in unpredictable order, results in far transfer to *Story Memory* recall, the measure of episodic memory that was found to be a significant predictor in the current study. Taken together, these results suggest that neuro-cognitive predictors of task learning can be informative about whether we can see potential transfer to tasks that have the same neuro-cognitive underpinnings. Another significance of the current study is the potential identification of individuals who may benefit most from BWGU training. Episodic memory is considered to be an early marker of mild cognitive impairment; therefore, it is possible that BWGU training may be beneficial to not only healthy older adults but to build a reserve in a key cognitive function known to be impacted in at-risk older adults, such as those with mild cognitive impairment. Finally, forecasting analysis on the time series of the game shows that day-to-day psychosocial wellbeing and hours of sleep can impact the game performance of that day or of the next day, but only in about 50% of participants in this study. Others did not exhibit any relationship between these daily measures (sleep and wellbeing) and game performance. Large-scale studies are warranted to understand why some older adults show such dependencies, and whether resistance to such dependencies results in the long-term maintenance of cognition. Importantly, data from these time series forecasting suggest that for a large proportion of individuals, the efficacy of the intervention can be improved at an individual level by incorporating sleep and psychosocial factors into the closed-loop individualized-adaptive feedback design. Identification through such modeling of how these individual-level daily variables (task performance, sleep, mood, etc.) impact our learning during an intervention can help us develop more efficacious, individualized cognitive interventions.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories

and accession number(s) can be found at: https://utdallas.box.com/s/c50xa6jg7u28kmyume070s4jitreecsu.

Ethics statement

The studies involving human participants were reviewed and approved by University of Texas at Dallas Institutional Review Board. The participants provided their written informed consent to participate in this study.

Author contributions

CB designed the original BirdWatch Game in Matlab and the cognitive intervention. The game was further developed by CB, ES, and PF as BWGU in the Unity platform. CB developed the study as a Principal Investigator, with the assistance of PF and DP, who were co-investigators on the project, evaluated individual subjects learning rates across various functions, conducted analyses of BWGU arm differences, and edited and co-wrote the manuscript. ES, PS, AS, and SQ collected the data for the clinical trial. PS and SQ performed the MRI preprocessing and FreeSurfer analysis of the structural data under CB's direction. ES performed all other analyses and wrote the first version of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by a grant from the National Institute on Aging to CB (titled Strategic Training to Optimize Neurocognitive Functions in Older Adults under Award R56AG060052, PI: CB).

Acknowledgments

The authors acknowledge Eric Shad Miller for his contributions to implementing the BWGU paradigm in the Unity game engine. We thank Glenn Hulon Sherard and Kristen Platt for assistance with behavioral data collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.936528/full#supplementary-material

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *J. Exp. Psychol.* 117, 288–318. doi: 10.1037/0096-3445.117.3.288

Adólfsdóttir, S., Haász, J., Wehling, E., Ystad, M., Lundervold, A., and Lundervold, A. J. (2014). Salient measures of inhibition and switching are associated with frontal lobe gray matter volume in healthy middle-aged and older adults. *Neuropsychology* 28, 859–869. doi: 10.1037/neu0000082

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. Budapest: Academiai Kaido.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* 52, 317–332. doi: 10.1007/BF02294359

Badre, D., and Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia* 45, 2883–2901. doi: 10.1016/j.neuropsychologia.2007.06.015

Bak, T. H., Long, M. R., Vega-Mendoza, M., and Sorace, A. (2016). Novelty, challenge, and practice: the impact of intensive language learning on attentional functions. *PLoS ONE* 11, e0153485. doi: 10.1371/journal.pone.0153485

Basak, C., Boot, W. R., Voss, M. W., and Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychol. Aging* 23, 765–777. doi: 10.1037/a0013494

Basak, C., and O'Connell, M. A. (2016). To switch or not to switch: role of cognitive control in working memory training in older adults. *Front. Psychol.* 7, 230. doi: 10.3389/fpsyg.2016.00230

Basak, C., Qin, S., and O'Connell, M. A. (2020). Differential effects of cognitive training modules in healthy aging and mild cognitive impairment: a comprehensive meta-analysis of randomized controlled trials. *Psychol. Aging* 35, 220–249. doi: 10.1037/pag0000442

Basak, C., and Verhaeghen, P. (2011). Aging and switching the focus of attention in working memory: age differences in item availability but not in item accessibility. *J. Gerontol. Ser. B* 66B, 519–526. doi: 10.1093/geronb/gbr028

Basak, C., Voss, M. W., Erickson, K. I., Boot, W. R., and Kramer, A. F. (2011). Regional differences in brain volume predict the acquisition of skill in a complex real-time strategy videogame. *Brain Cogn.* 76, 407–414. doi: 10.1016/j.bandc.2011.03.017

Bean, J. (2011). "Rey auditory verbal learning test, rey AVLT," in *Encyclopedia of Clinical Neuropsychology*, eds J. S. Kreutzer, J. DeLuca, and B. Caplan (New York, NY: Springer), 2174–2175. doi: 10.1007/978-0-387-79948-3_1153

Beaunieux, H., Hubert, V., Pitel, A. L., Desgranges, B., and Eustache, F. (2009). Episodic memory deficits slow down the dynamics of cognitive procedural learning in normal ageing. *Memory* 17, 197–207. doi: 10.1080/09658210802212010

Beaunieux, H., Hubert, V., Witkowski, T., Pitel, A. -L., Rossi, S., Danion, J. -M., et al. (2006). Which processes are involved in cognitive procedural learning? *Memory* 14, 521–539. doi: 10.1080/09658210500477766

Boot, W. R., Basak, C., Erickson, K. I., Neider, M., Simons, D. J., Fabiani, M., et al. (2010). Transfer of skill engendered by complex task training under conditions of variable priority. *Acta Psychol.* 135, 349–357. doi: 10.1016/j.actpsy.2010.09.005

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *J. Math. Psychol.* 44, 62–91. doi: 10.1006/jmps.1999.1277

Brehmer, Y., Westerberg, H., and Bäckman, L. (2012). Working-memory training in younger and older adults: training gains, transfer, and maintenance. *Front. Hum. Neurosci.* 6, 63. doi: 10.3389/fnhum.2012.00063

Brown, E. M., Pierce, M. E., Clark, D. C., Fischl, B. R., Iglesias, J. E., Milberg, W. P., et al. (2020). Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *NeuroImage* 210, 116563. doi: 10.1016/j.neuroimage.2020.116563

Burgess, N., Maguire, E. A., and O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron* 35, 625–641. doi: 10.1016/S0896-6273(02)00830-9

Bürki, C. N., Ludwig, C., Chicherio, C., and de Ribaupierre, A. (2014). Individual differences in cognitive plasticity: an investigation of training curves in younger and older adults. *Psychol. Res.* 78, 821–835. doi: 10.1007/s00426-014-0559-3

Cuenen, A., Jongen, E. M. M., Brijs, T., Brijs, K., Houben, K., and Wets, G. (2016). Effect of a working memory training on aspects of cognitive ability and driving ability of older drivers: merits of an adaptive training over a non-adaptive training. *Transport. Res. F* 42, 15–27. doi: 10.1016/j.trf.2016.06.012

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Doppler, C. E. J., Meyer, L., Dovern, A., Stühmer-Beckh, J., Weiss, P. H., and Fink, G. R. (2019). Differential impact of social and monetary reward on procedural learning and consolidation in aging and its structural correlates. *Front. Aging Neurosci.* 11, 188. doi: 10.3389/fnagi.2019.00188

Drozdick, L. W., Wahlstrom, D., Zhu, J., and Weiss, L. G. (2012). "The Wechsler adult intelligence scale-fourth edition and the wechsler memory scale," in *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, 3rd Edn, eds D. P. Flanagan and P. L. Harrison (New York, NY: The Guilford Press), 197–223.

Erickson, K. I., Boot, W. R., Basak, C., Neider, M. B., Prakash, R. S., Voss, M. W., et al. (2010). Striatal volume predicts level of video game skill acquisition. *Cerebr. Cortex* 20, 2522–2530. doi: 10.1093/cercor/bhp293

Fedorenko, E., and Thompson-Schill, S. L. (2014). Reworking the language network. *Trends Cogn. Sci.* 18, 120–126. doi: 10.1016/j.tics.2013.12.006

Festini, S. B., McDonough, I. M., and Park, D. C. (2016). The busier the better: greater busyness is associated with better cognition. *Front. Aging Neurosci.* 8, 98. doi: 10.3389/fnagi.2016.00098

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Gathercole, S. E., Dunning, D. L., Holmes, J., and Norris, D. (2019). Working memory training involves learning new skills. *J. Mem. Lang.* 105, 19–42. doi: 10.1016/j.jml.2018.10.003

Hagoort, P. (2013). MUC (Memory, Unification, Control) and beyond. *Front. Psychol.* 4, 416. doi: 10.3389/fpsyg.2013.00416

Holanda Júnior, F. W. N., and de Almondes, K. M. (2016). Sleep and executive functions in older adults: a systematic review. *Dement. Neuropsychol.* 10, 185–197. doi: 10.1590/S1980-5764-2016DN1003004

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2021). *Forecast: Forecasting Functions for Time Series and Linear Models. R Package Version 8.14.* Available online at: https://pkg.robjhyndman.com/forecast (accessed June 21, 2021).

Hyndman, R. J., and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *J. Statist. Softw.* 26, 1–22. doi: 10.18637/jss.v027.i03

Ihle, A., Ghisletta, P., Gouveia, É. R., Gouveia, B. R., Oris, M., Maurer, J., et al. (2021). Lower executive functioning predicts steeper subsequent decline in well-being only in young-old but not old-old age. *Int. J. Behav. Dev.* 45, 97–108. doi: 10.1177/0165025420937076

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., and Meier, B. (2010). The concurrent validity of the N -back task as a working memory measure. *Memory* 18, 394–412. doi: 10.1080/09658211003702171

Lampit, A., Hallock, H., and Valenzuela, M. (2014). Computerized cognitive training in cognitively healthy older adults: a systematic review and meta-analysis of effect modifiers. *PLoS Med.* 11, e1001756. doi: 10.1371/journal.pmed.1001756

Lim, Y. Y., Baker, J. E., Bruns, L., Mills, A., Fowler, C., Fripp, J., et al. (2020). Association of deficits in short-term learning and Aβ and hippocampal volume in cognitively normal adults. *Neurology* 95, e2577–e2585. doi: 10.1212/WNL.0000000000010728

Lo, J. C., Groeger, J. A., Cheng, G. H., Dijk, D. -J., and Chee, M. W. L. (2016). Self-reported sleep duration and cognitive performance in older adults: a systematic review and meta-analysis. *Sleep Med.* 17, 87–98. doi: 10.1016/j.sleep.2015.08.021

Lojo-Seoane, C., Facal, D., Guàrdia-Olmos, J., Pereiro, A. X., Campos-Magdaleno, M., Mallo, S. C., et al. (2020). Cognitive reserve and working memory in cognitive performance of adults with subjective cognitive complaints: longitudinal structural equation modeling. *Int. Psychogeriatr.* 32, 515–524. doi: 10.1017/S1041610219001248

López-Higes, R., Prados, J. M., Rubio-Valdehita, S., Rodríguez-Rojo, I., de Frutos-Lucas, J., Montenegro, M., et al. (2018). Factors explaining language performance after training in elders with and without subjective cognitive decline. *Front. Aging Neurosci.* 10, 264. doi: 10.3389/fnagi.2018.00264

Luerssen, A., and Ayduk, O. (2017). "Executive functions promote well-being: outcomes and mediators," in *The Happy Mind: Cognitive Contributions to Well-Being*, eds M. D. Robinson and M. Eid (Berlin: Springer International Publishing), 59–75. doi: 10.1007/978-3-319-58763-9_4

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide, 2nd Edn.* (London: Lawrence Erlbaum Associates Publishers), 492.

Maiovis, P., Ioannidis, P., Nucci, M., Gotzamani-Psarrakou, A., and Karacostas, D. (2016). Adaptation of the Cognitive Reserve Index Questionnaire (CRIq) for the Greek population. *Neurol. Sci.* 37, 633–636. doi: 10.1007/s10072-015-2457-x

Mihalca, L., Salden, R. J. C. M., Corbalan, G., Paas, F., and Miclea, M. (2011). Effectiveness of cognitive-load based adaptive instruction in genetics education. *Comput. Hum. Behav.* 27, 82–88. doi: 10.1016/j.chb.2010.05.027

Miyake, A., and Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: four general conclusions. *Curr. Direct. Psychol. Sci.* 21, 8–14. doi: 10.1177/0963721411429458

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Newson, R. S., and Kemps, E. B. (2006). The nature of subjective cognitive complaints of older adults. *Int. J. Aging Hum. Dev.* 63, 139–151. doi: 10.2190/1EAP-FE20-PDWY-M6P1

Nucci, M., Mapelli, D., and Mondini, S. (2012). Cognitive Reserve Index questionnaire (CRIq): a new instrument for measuring cognitive reserve. *Aging Clin. Exp. Res.* 2012, 24. doi: 10.1037/t53917-000

Oberauer, K. (2006). Is the focus of attention in working memory expanded through practice? *J. Exp. Psychol.* 32, 197–214. doi: 10.1037/0278-7393.32.2.197

O'Connell, M. A., and Basak, C. (2018). Effects of task complexity and age-differences on task-related functional connectivity of attentional networks. *Neuropsychologia* 114, 50–64. doi: 10.1016/j.neuropsychologia.2018.04.013

Opdebeeck, C., Martyr, A., and Clare, L. (2016). Cognitive reserve and cognitive function in healthy older people: a meta-analysis. *Neuropsychol. Dev. Cogn.* 23, 40–60. doi: 10.1080/13825585.2015.1041450

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59. doi: 10.1002/hbm.20131

Ozakbas, S., Yigit, P., Akyuz, Z., Sagici, O., Abasiyanik, Z., Ozdogar, A. T., et al. (2021). Validity and reliability of "Cognitive Reserve Index Questionnaire" for the Turkish Population. *Multiple Scler. Relat. Disord.* 50, 102817. doi: 10.1016/j.msard.2021.102817

Papoulis, A. (2002). *Probability, Random Variables, and Stochastic Processes*. New York, NY: Tata McGraw-Hill Education.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., and Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychol. Aging* 17, 299–320. doi: 10.1037/0882-7974.17.2.299

Park, D. C., Lodi-Smith, J., Drew, L., Haber, S., Hebrank, A., Bischof, G. N., et al. (2014). The impact of sustained engagement on cognitive function in older adults: The synapse project. *Psychol. Sci.* 25, 103–112. doi: 10.1177/0956797613499592

Payne, B. R., Jackson, J. J., Noh, S. R., and Stine-Morrow, E. A. L. (2011). In the zone: flow state and cognition in older adults. *Psychol. Aging* 26, 738–743. doi: 10.1037/a0022359

Plieger, T., and Reuter, M. (2020). Stress and executive functioning: a review considering moderating factors. *Neurobiol. Learn. Mem.* 173, 107254. doi: 10.1016/j.nlm.2020.107254

Posner, M. I. (1980). Orienting of attention. *Quart. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Qin, S., and Basak, C. (2020). Age-related differences in brain activation during working memory updating: an fMRI study. *Neuropsychologia* 138, 107335. doi: 10.1016/j.neuropsychologia.2020.107335

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: http://www.R-project.org/

Rana, B. K., Panizzon, M. S., Franz, C. E., Spoon, K. M., Jacobson, K. C., Xian, H., et al. (2018). Association of sleep quality on memory-related executive functions in middle age. *J. Int. Neuropsychol. Soc.* 24, 67–76. doi: 10.1017/S1355617717000637

Ray, N. R., O'Connell, M. A., Nashiro, K., Smith, E. T., Qin, S., and Basak, C. (2017). Evaluating the relationship between white matter integrity, cognition, and varieties of video game learning. *Restorat. Neurol. Neurosci.* 35, 437–456. doi: 10.3233/RNN-160716

Rozas, A. X. P., Juncos-Rabadán, O., and González, M. S. R. (2008). Processing speed, inhibitory control, and working memory: three important factors to account for age-related cognitive decline. *Int. J. Aging Hum. Dev.* 66, 115–130. doi: 10.2190/AG.66.2.b

Saint-Cyr, J. A., and Taylor, A. E. (1992). "The mobilization of procedural learning: the "key signature" of the basal ganglia," in *Neuropsychology of Memory, 2nd Edn*, eds L. R. Squire and N. Butters (New York, NY: Guilford Press), 188–202.

Saint-Cyr, J. A., Taylor, A. E., and Lang, A. E. (1988). Procedural learning and neostriatal dysfunction in man. *Brain* 111, 941–959.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* 46, 1273. doi: 10.2307/1913828

Shields, G. S., Sazma, M. A., and Yonelinas, A. P. (2016). The effects of acute stress on core executive functions: a meta-analysis and comparison with cortisol. *Neurosci. Biobehav. Rev.* 68, 651–668. doi: 10.1016/j.neubiorev.2016.06.038

Simonyan, K. (2019). Recent advances in understanding the role of the basal ganglia. *F1000Research* 8, 122. doi: 10.12688/f1000research.16524.1

Smith, E. T., Bhaskar, B., Hinerman, A., and Basak, C. (2020). Past gaming experience and cognition as selective predictors of novel game learning across different gaming genres. *Front. Psychol.* 11, 786. doi: 10.3389/fpsyg.2020.00786

Srinivasan, D., Erus, G., Doshi, J., Wolk, D. A., Shou, H., Habes, M., et al. (2020). A comparison of Freesurfer and multi-atlas MUSE for brain anatomy segmentation: findings about size and age bias, and inter-scanner stability in multi-site aging studies. *NeuroImage* 223, 117248. doi: 10.1016/j.neuroimage.2020.117248

Stawski, R. S., Mogle, J., and Sliwinski, M. J. (2011). Intraindividual coupling of daily stressors and cognitive interference in old age. *J. Gerontol. Ser. B* 66B(Suppl.1), i121–i129. doi: 10.1093/geronb/gbr012

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychol. Rev.* 120, 439–471. doi: 10.1037/a0033138

Tucker, A. M., and Stern, Y. (2011). Cognitive reserve in aging. *Curr. Alzheimer's Res.* 8, 354–360. doi: 10.2174/156720511795745320

Verhaeghen, P., Cerella, J., and Basak, C. (2004). A working memory workout: how to expand the focus of serial attention from one to four items in 10 hours or less. *J. Exp. Psychol.* 30, 1322–1337. doi: 10.1037/0278-7393.30.6.1322

Ward, A., Alberg Sorensen, K., Kousgaard, H., Schack Thoft, D., and Parkes, J. (2020). Going back to school – an opportunity for lifelong learning for people with dementia in Denmark (Innovative practice). *Dementia* 19, 2461–2468. doi: 10.1177/1471301218763190

Zavecz, Z., Nagy, T., Galkó, A., Nemeth, D., and Janacsek, K. (2020). The relationship between subjective sleep quality and cognitive performance in healthy young adults: evidence from three empirical studies. *Sci. Rep.* 10, 4855. doi: 10.1038/s41598-020-61627-6

Zhu, J., Zhu, D., Zhang, C., Wang, Y., Yang, Y., and Yu, Y. (2019). Quantitative prediction of individual cognitive flexibility using structural MRI. *Brain Imag. Behav.* 13, 781–788. doi: 10.1007/s11682-018-9905-1

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fnagi.2022.897343

OPEN ACCESS

EDITED BY
William Kremen,
University of California, San Diego,
United States

REVIEWED BY
Karl Healey,
Michigan State University,
United States
Sheila Black,
University of Alabama, United States

*CORRESPONDENCE
Zita Oravecz
zita@psu.edu

SPECIALTY SECTION
This article was submitted to Neurocognitive Aging and Behavior, a section of the journal Frontiers in Aging Neuroscience

RECEIVED 16 March 2022
ACCEPTED 23 August 2022
PUBLISHED 26 September 2022

CITATION
Oravecz Z, Harrington KD, Hakun JG, Katz MJ, Wang C, Zhaoyang R and Sliwinski MJ (2022) Accounting for retest effects in cognitive testing with the Bayesian double exponential model via intensive measurement burst designs.
Front. Aging Neurosci. 14:897343.
doi: 10.3389/fnagi.2022.897343

Accounting for retest effects in cognitive testing with the Bayesian double exponential model via intensive measurement burst designs

Zita Oravecz[1,2,3]*, Karra D. Harrington[3], Jonathan G. Hakun[3,4,5], Mindy J. Katz[6], Cuiling Wang[7], Ruixue Zhaoyang[3] and Martin J. Sliwinski[1,3]

[1]Department of Human Development and Family Studies, Pennsylvania State University, University Park, PA, United States, [2]Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA, United States, [3]Center for Healthy Aging, Pennsylvania State University, University Park, PA, United States, [4]Department of Neurology, Pennsylvania State University, Hershey, PA, United States, [5]Department of Psychology, Pennsylvania State University, University Park, PA, United States, [6]Department of Neurology, Albert Einstein College of Medicine, Bronx, NY, United States, [7]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, United States

Monitoring early changes in cognitive performance is useful for studying cognitive aging as well as for detecting early markers of neurodegenerative diseases. Repeated evaluation of cognition via a measurement burst design can accomplish this goal. In such design participants complete brief evaluations of cognition, multiple times per day for several days, and ideally, repeat the process once or twice a year. However, long-term cognitive change in such repeated assessments can be masked by short-term within-person variability and retest learning (practice) effects. In this paper, we show how a Bayesian double exponential model can account for retest gains across measurement bursts, as well as warm-up effects within a burst, while quantifying change across bursts in peak performance. We also highlight how this approach allows for the inclusion of person-level predictors and draw intuitive inferences on cognitive change with Bayesian posterior probabilities. We use older adults' performance on cognitive tasks of processing speed and spatial working memory to demonstrate how individual differences in peak performance and change can be related to predictors of aging such as biological age and mild cognitive impairment status.

KEYWORDS

retest learning, measurement burst design, double negative exponential model, subtle cognitive decline, Bayesian multilevel modeling

Introduction

Accurate and sensitive measurement of cognitive change is required to advance the understanding of normative cognitive aging and improve the detection of the subtle cognitive changes that are associated with the preclinical stages of neurodegenerative diseases, such as Alzheimer's disease. Although cumulative cognitive change over the course of decades is quite robust, the amount of change expected over a year or two is quite subtle, even in the case of prodromal disease (Baker et al., 2016). Traditional methods relying on infrequent lab-based assessment of cognitive performance make it difficult to differentiate changes due to cognitive aging, progression of neurodegenerative disease, and the possible effects of interventions designed to improve or slow decline in cognitive function. A major challenge to detecting subtle cognitive change is the presence of retest, or practice effects, which refer to the ubiquitous finding that performance on cognitive tests improves with repeated testing.

Although widely recognized as biasing longitudinal estimates and intervention effects, there is no consensus on best methods to address retest effects (Jones, 2015). Indeed, it is extremely difficult to disentangle retest related effects from other sources of change (e.g., aging, disease progression, and interventions) using data from conventional longitudinal designs that consist of repeated single-shot assessments, usually spaced over long time intervals. **Figure 1** illustrates this point by depicting hypothetical longitudinal data in which observed performance (black line) reflects a mixture of two latent processes, retest related gains (red line) and aging-related declines (blue line). Panel A shows a case in which change in performance is flat, where the stability is a product of retest related gains offsetting aging-related decline. Comparing manifest performance (black lines) in **Figure 1**, panels B and C suggests that the former exhibited more cognitive decline and that neither exhibited evidence of improvement in cognitive performance that could be due to retest effects. However, the underlying latent aging effects show equivalent longitudinal decrements in Panels B and C, but differential latent retest effects.

This example illustrates two important points. First, processes that drive retest effects may be operating even if manifest performance shows no improvement or even a decline. That is, one cannot take the absence of overt performance gains as evidence that retest effects are absent. Moreover, even in the presence of manifest decline in cognitive performance, retest effects may be a significant confound that obscures important individual differences. Second, failure to accurately characterize and account for retest gains could add considerable noise and bias when testing for the effects of interventions, biological markers of aging or disease progression, or other exposures (e.g., stress, environmental toxicants) on cognitive trajectories.

Conventional longitudinal designs place significant constraints on approaches for disentangling retest effects from other types of change. The use of a control group which receives their first exposure to a cognitive test at follow-up may be useful for estimating bias in the group averages but cannot assist in correcting for retest effects at the individual level. Statistical control procedures that involve covarying for the number of retest assessments are susceptible to bias and are especially sensitive to assumptions regarding the presence of age-cohort effects (Hoffman et al., 2011). To overcome these limitations, our approach utilizes a measurement burst design (Sliwinski, 2008) which consists of closely spaced "bursts" of repeated measurements which are repeated over longer intervals. This type of intensive longitudinal design (ILD) permits modeling of retest effects using repeated administrations over a short interval within bursts (e.g., daily) to render long-term retest effects negligible and to model long-term trends using measurements bursts repeated over longer intervals (e.g., annually) across bursts (Sliwinski et al., 2010; Rast et al., 2012). In a proof of concept, Munoz et al. (2015) fit a non-linear multilevel model to measurement burst data to disentangle short-term retest effects from long-term declines in asymptotic performance.

We propose a psychometric cognitive process model, the Bayesian double exponential model (BDEM) to disentangle retest learning effects from longitudinal changes in asymptotic performance. The BDEM allows parameterizing performance in terms of distinct retest features including learning rate (how quickly someone reaches peak performance), retest gains (how much overall improvement is observed), peak (asymptotic) performance, and warm-up effects that occur at the beginning of follow-up bursts. Once practice effects are accounted for, we can link individual differences in peak performance and changes in peak performance to person-level indicators such as age and mild cognitive impairment (MCI) status.

While the primary aim of BDEM is to disentangle learning features from peak performance (with the goal of modeling asymptotic change over time), each model parameter may also be of interest for understanding the dynamics of cognitive change. For this reason, we also quantify individual differences, in a multilevel framework, not only in terms of peak performance and changes therein, but also for example in terms of learning rate, and intra-individual variability in performance, and test whether these are linked to cognitive aging (Lövdén et al., 2007) or MCI status (Cerino et al., 2021).

Compared to earlier work with the double negative exponential model, such as in Sliwinski et al. (2010), Broitman et al. (2020), our approach casts the double negative exponential model in a multilevel Bayesian statistical framework, which has two main advantages. First, it allows for all double exponential parameters to be person-specific and be regressed on person-level predictors in a single step analysis, this way improving estimation accuracy. Second, it allows for a more nuanced

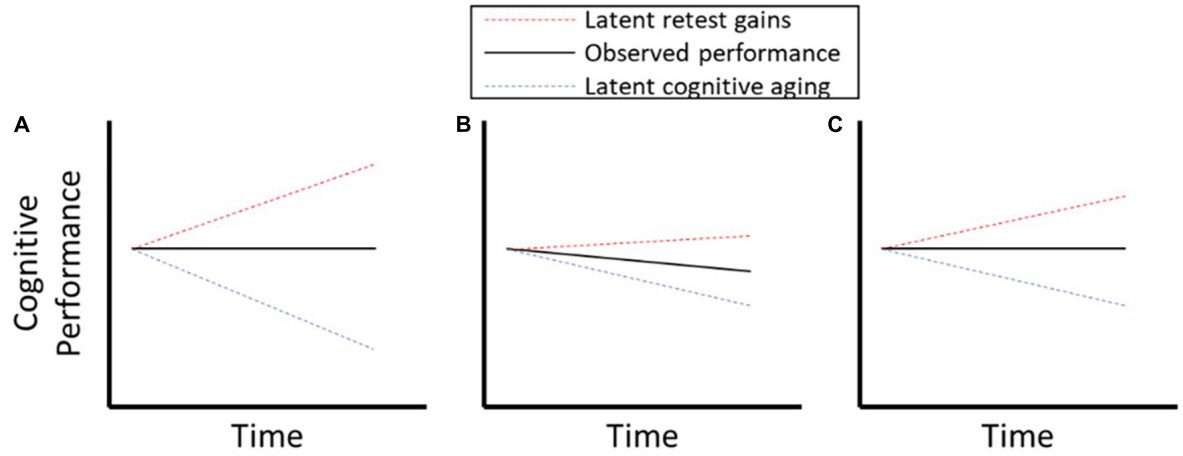


FIGURE 1
(A–C) Illustration of retest effects confounding measurement of cognitive decline.

inference in terms of person-specific characteristics, for example the risk of cognitive decline can be articulated in terms of individual specific probabilities, as illustrated later in the paper.

In the current study we analyzed data from the Einstein Aging Study (EAS; Zhaoyang et al., 2021; Katz et al., 2021), a longitudinal study that included annual conventional assessments and ambulatory assessment bursts in a racially diverse, systematically recruited community dwelling cohort of older adults (age 70+). We evaluated the descriptive adequacy of the BDEM to EAS data obtained from high frequency cognitive assessments completed by participants using mobile devices in naturalistic settings. We also examined whether BDEM parameters, such as asymptotic performance, change in asymptotic performance, learning rate, and intra-individual variability, differentiated among individuals across different ages and MCI status.

Materials and methods

Study design and procedure

Data were drawn from the ongoing EAS, a prospective, longitudinal study of risk factors for MCI and dementia. Systematic random sampling from New York City Registered Voter Lists for Bronx County was used to recruit participants. Further screening of participants was conducted via telephone to ensure that participants met the study inclusion criteria: English-speaking, community-residing, ambulatory, and aged over 70 years. Exclusion criteria were: significant hearing or vision loss, current substance abuse, severe psychiatric symptoms, chronic medicinal use of opioids or glucocorticoids, treatment for cancer within the past 12 months, and diagnosis of dementia. All participants provided written informed consent and the study was approved by the Albert Einstein College of Medicine Institutional Review Board.

Figure 2 shows an illustration of the overall measurement burst design deployed in the EAS project. Each year participants completed a combination of clinic-based assessments and ambulatory ecological momentary assessments (EMA). After telephone screening, eligible participants were invited to attend two in-person clinic-based assessments. The first assessment day included completing self-report questionnaires and neuropsychological assessment. The second assessment day included a 1.5-h training session on how to use the study-provided smartphone and complete the EMA portion of the study. Participants who were assessed between March and June 2020 completed modified versions of these assessments and training remotely via telephone and received the study smartphone via a package delivery service.

The ambulatory burst component of the study took place in participants' natural environments. While participants went about their daily activities, they completed six brief assessments (up to 5 mins each) during their typical waking hours, over a period of 16th days—these assessments together formed a "burst." The assessments included brief self-report questions as well as the cognitive assessments. The protocol included a self-initiated wake-up assessment, a self-initiated end-of-day assessment, and four quasi-random "beeped" assessments that participants received a notification from the study phone to complete. The beeped assessments were schedule approximately 3.5 h apart, and times varied across the days of the week. After the ambulatory burst period, participants returned the study smartphone at a third clinic visit and the data were downloaded from the phone.

In the present study we analyzed baseline demographic and MCI status data, as well as cognitive performance data from Burst 1 and Burst 2, that were collected between May 2017

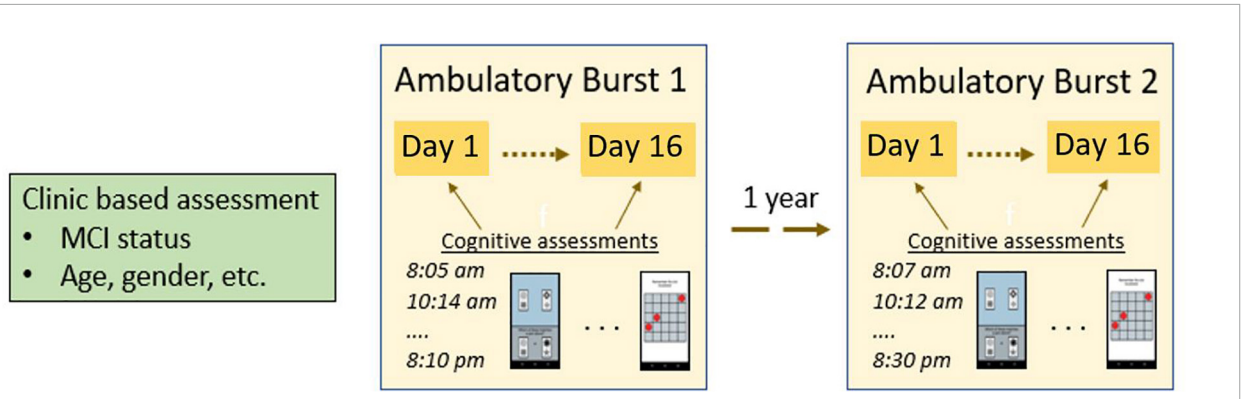


FIGURE 2
Illustration of a measurement burst design with two bursts.

and June 2020. The sample consisted of 318 adults, of which 53.8% (n = 171) completed both bursts, while the remaining participants had only baseline (Burst 1) data. Of the 147 participants who did not have follow up (Burst 2) data, 31.3% (n = 46) had not yet been contacted for follow up, 24.0% (n = 35) had chosen to not complete the EMA component of the study, 4.0% (n = 6) had missing or unusable data on the smartphone, 8.2% (n = 12) were unable to participate due to illness or death, and 36.7% (n = 54) were withdrawn. Characteristics of the sample are provided in the "Results" section.

Measures

Demographics

Participants self-reported demographic details via questionnaire, including age in years, sex (male/female), race and ethnicity (White/Black/Hispanic White/Hispanic Black/Asian/more than one race), and education (years in school).

Mild cognitive impairment status

As part of their participation all participants underwent neuropsychological assessment to determine their cognitive status. The neuropsychological assessment included measures of memory, executive function, attention, language, and visuospatial ability. Free recall from the Free and Cued Selective Reminding Test (Buschke, 1984) and delayed recall of the Benson Complex Figure (Possin et al., 2011) assessed memory function; Trail Making Test – Part B (Reitan, 1958) and Phonemic Verbal Fluency (Tombaugh et al., 1999) assessed executive function; Trail Making Test – Part A (Reitan, 1958) and WAIS-III Digit Span (Wechsler, 1987) assessed attention; Multilingual Naming Test (Ivanova et al., 2013) and Category Fluency (Monsch et al., 1992) assessed language; and immediate recall of Benson Complex Figure (Possin et al., 2011) and WAIT-III Block Design (Wechsler, 1987) assessed visuospatial function. MCI status was classified algorithmically using criteria from described in Jak et al. (2009) and described in detail in Katz et al. (2021). Briefly, criteria included: (a) impaired scores on two measures of the same cognitive domain; or (b) one impaired score in three out of five cognitive domains; or (c) having functional decline assessed by the Instrumental Activities of Daily Living Scale (Lawton and Brody, 1969). Impairment was defined as scores >1 SD below age-, sex-, and education-adjusted normative mean.

Symbol search task

The symbol search task, shown on the left side of **Figure 3**, measures processing speed. In the current study, on each trial of the task, participants saw three symbol pairs at the top of the screen and two symbol pairs at the bottom of the screen. They were instructed to match as quickly and accurately as they could one of the two pairs presented at the bottom to one of the three pairs at the top. Participants completed 11 trials per session. We analyzed daily aggregates of response times on correct trials with the BDEM.

Grid memory task

Grid Memory is a free recall paradigm that assesses spatial working memory, shown on the right side of **Figure 3**. This task in the current study involved a brief study phase, during which 3 dots are presented at random locations on a 5 × 5 grid for 3 s, an 8-s letter-cancelation distractor phase, followed by free recall of locations occupied by dots during the study phase. The free recall phase required participants to touch the locations in an empty 5 × 5 grid where the 3 dots were presented initially. Participants completed two trials that incorporated all three phases per session. The outcome of interest for this task was the Euclidean distance between the location of the incorrect dot to the correct grid (0 if correct). This gave partial credit based on the deviation of the recalled compared to the correct dot locations. We refer to this error distance as "units of error" from here on.

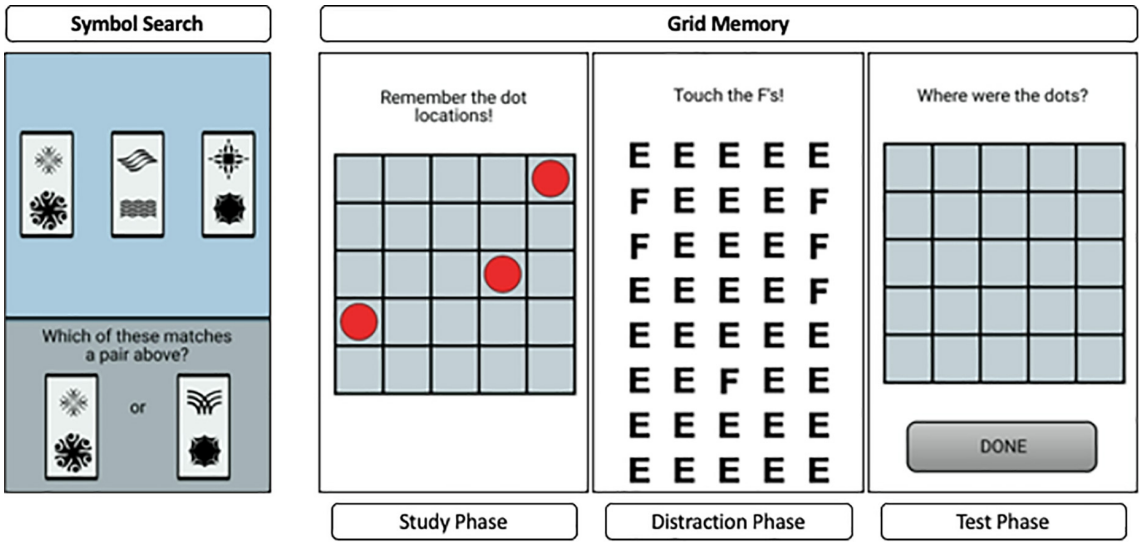


FIGURE 3
An example trial from the symbol search task **(top)** and the grid memory task **(bottom)**.

Data analysis with the double exponential model

First, we start by specifying a negative exponential model for repeatedly administered cognitive performance data close in time. This model can characterize change in performance in terms of four parameters: learning rate, retest gain, peak performance and intra-individual variability. By disentangling the latent processes in observed performance, the negative exponential model separates *how much* learning occurs (retest gain) from *how fast* the learning occurs (learning rate). The learning curve is characterized by an exponential shape, which is supported by studies on learning (see, e.g., Heathcote et al., 2000), as well as studies on aging (see, e.g., Sliwinski et al., 2010). These curves will be derived for every person to accurately dissociate learning from the person-level overall peak performance.

Consider a study that only has one burst of measurements. The top row of **Figure 4** shows a graphical representation of the negative exponential model (solid line) fit to a sequence of a simulated cognitive performance measure (indicated by dots), which in our case was either error distance or reaction time. We will refer to this participant as the "baseline" for later comparisons. At the start of the burst, their errors are distinctly higher than near the end; that is, the participant shows evidence of learning across sessions in a measurement burst. This improvement is modeled through a negative exponential function, which is parametrized as follows:

$$Y_{ti} = a_i + g_i \times \exp\left[-r_i \times M_{ti}\right] + e_{ti} \quad (1)$$

The cognitive performance data over sessions t from an individual i is denoted as Y_{ti}. Parameter a_i stands for the person's asymptotic or peak performance (best performance given unlimited practice), which was set to 2.00 in the current example shown in the top row of **Figure 4**. The gain in performance across measurements is quantified by g_i, which roughly corresponds to the height of the exponential (set to 0.50 in this example). The learning rate is captured by r_i (set to 0.30), the steepness of the exponential curve across measurements (with measurement occasions denoted as M_{ti}) in the study. Finally, e_{ti} is a time-and-person-specific error term, with mean zero and standard deviation $\sigma_{e,i}$ (set to 0.05), where $\sigma_{e,i}$ captures the within-person variations (i.e., intraindividual variability) across trials.

The bottom two rows of **Figure 4** shows four additional synthetic persons' data and model fit, each with one parameter different from the "baseline" in the top row. The participant in the left panel of the second row has better peak performance (less error): their asymptote settles at 1.75 instead of 2.00. The right panel of the second row shows a participant with a higher gain parameter across trials compared to the baseline person in the top row (g_i is 1.00 instead of 0.50); the exponential starts out higher compared to baseline. The bottom left panel depicts a faster learning rate (steeper exponential slope; r_i is set to 0.60 instead of 0.30) than the baseline; and it reaches peak performance faster (given the same amount of gain). Finally, the bottom right panel shows a participant with greater intra-individual variation ($\sigma_{e,i}$), as indicated by the more scattered dots around the fitted model (set to 0.10 instead of 0.05). Simulation analyses showed good recovery of these parameters with 10 data points per person, in terms of at least 95% of the

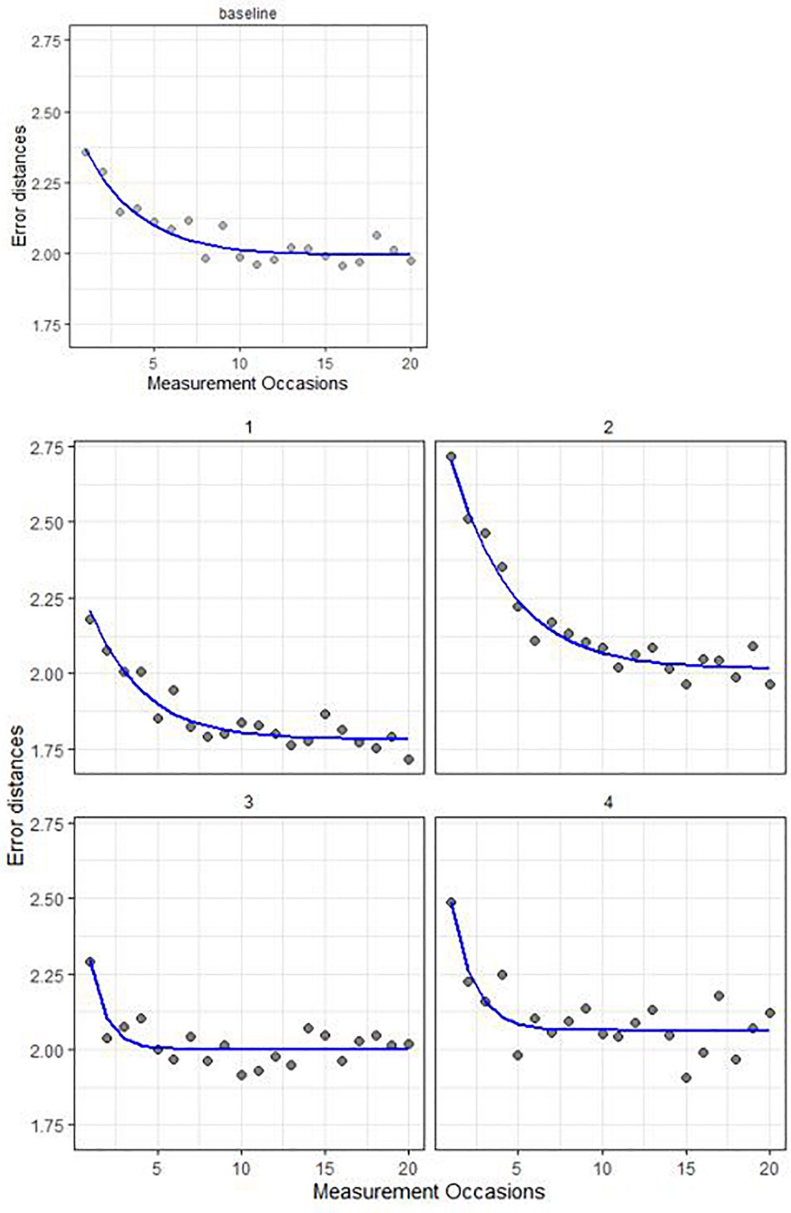


FIGURE 4
Five synthetic participant's data (gray dots) and model fit (solid line).

simulated values falling in the estimated 95% credible interval of their corresponding parameter estimates.

The double negative exponential model extends the previously introduced negative exponential model by considering retest gains across bursts as well. It is specified as:

$$Y_{ti} = g_i \times \exp[-r_i \times M_{ti}] + I(B_{ti} > 1) \times g_i^* \times \exp[-r_i \times T_{ti}] + e_{ti} + a_i + \triangle_i \times (B_{ti} - 1) \quad (2)$$

Parameters g_i and r_i again represent gain and learning rate, as in Eq. (1), but now we have two sets of them: one (g_i, r_i) set to capture *continuous learning throughout the study* [much like in Eq. (1) when we only had one burst], shown in the first line of the Eq. (2), and a second (g_i^*, r_i^*) set that represents *warm-up processes after the first burst* [i.e., $I(B_{ti} > 1)$] – with gain and learning rate parameters denoted by an asterisk (*), as shown in the second line of Eq. (2). This warm-up effect also has an exponential functional form, and it operates on the measurements nested within a burst (denoted with T_{ti}).

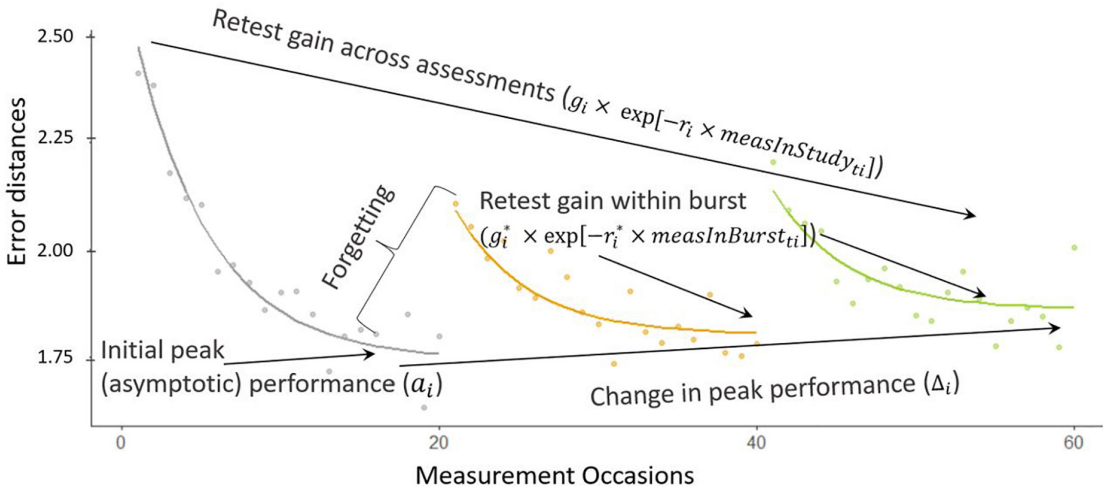


FIGURE 5
Illustration of the double negative exponential model.

Similarly to Eq. (1), e_{ti} again represents a time and person-specific error term, with its standard deviation $\sigma_{e,i}$ quantifying the intraindividual variability in performance across all trials.

Most importantly, as shown in $\Delta_i \times (B_{ti} - 1)$ of Eq. (2), we are now modeling the change in asymptotic (peak) performance between bursts. This is accomplished by parameter Δ_i, which adjusts the asymptotic performance (a_i), by some magnitude of change in every burst following the first one. We then investigate individual differences in these key parameters by adding covariates such as age, sex, and MCI status to the model.

Figure 5 shows a graphical representation of the double negative exponential model fit to cognitive performance measures (error distances in this example, displayed as dots) over trials t from a synthetic individual i, over three bursts (note that in the dataset analyzed later there are only two bursts, but we display three here for illustration purposes of the general approach). We can see that in the beginning of the study, there are more errors than at the end of the study, while also within each burst the first error rates are higher than the rest. We also observe retest gain across assessments: a learning process across the whole study period (here parameter g_i quantifies person i's gain across all measurements in the study, and r_i represents their corresponding learning rate). Also, in each burst after the first, there is a warm-up gain, a within-burst learning process with gain g_i^* and learning rate r_i^* parameters. In the first burst of data (over measurement occasions 1−20), the asymptote represents the person-specific initial peak performance, which becomes worse with every burst in this example (higher asymptotes correspond to more error). The amount of change in peak performance is quantified by Δ_i. As can also be seen in **Figure 5**, due to the retest gain across assessments and within burst, there seems to be an improvement in performance (decrease in error distances overall across the study). However, if we look at the long-term change in terms of the peak performance parameter of the proposed model (i.e., change in asymptote), there is in fact an incremental decline in performance, manifested through an increase in errors (i.e., worsening peak performance across bursts).

The warm-up effect represents an expected decrease in performance from the peak performance of a prior burst to the initial performance at its follow-up burst. It is an important process to model as this decrease may not reflect true "cognitive decline" and could instead represent some "forgetting" of testing procedures (see also in e.g., Dutilh et al., 2009). Our proposed modeling approach captures participants recovering their previous gains overlaid on their continuous improvement.

Modeling cognitive performance in terms of the person-specific double negative exponential parameters can help us capture retest effects and isolate them from other cognitive performance indicators. Investigating possible sources of individual differences (e.g., age, MCI status) in these cognitive parameters (i.e., learning rate, change in peak performance, etc.) can help us learn about processes related to retest effects and cognitive decline. In summation, our proposed model represents a cognitive psychometric approach to interpreting cognitive performance data. This model will require a nuanced and flexible statistical framework for inference. We chose a multilevel Bayesian framework (Gelman and Hill, 2007) for implementing the double negative exponential model, discussed next.

The multilevel specifications of the double exponential model

In our proposed modeling approach, all double negative exponential parameters, a_i, Δ_i, g_i, r_i, g_i^*, r_i^*, and $\sigma_{e,i}$ were

allowed to be person-specific and are pooled together via group-level (population) distribution. This represents a standard multilevel modeling approach (Raudenbush and Bryk, 2002) that has been proven useful for improving estimation accuracy, as it allows for the person- and group-level trends to simultaneously inform each other. Moreover, we also aim at identifying possible sources of individual differences in these parameters. Therefore, we regress person-specific cognitive performance characteristics (e.g., peak performance, change in peak performance, etc.) on a set of predictors, such as age and MCI status. Note that these cognitive performance characteristics are themselves model parameters, therefore they are estimated with error. We use a one-step approach to regress them on predictors to avoid generated regressor bias (Pagan, 1984).

More specifically, in our multilevel specification of the DNE, the peak performance, a_i, changes in peak performance, $\triangle_i$, and intra-individual variation in performance, $\sigma_{e,i}$, and learning rates (r_i, r_i^*) have group-level (population) distributions, the means of which are decomposed into products of predictors and regression coefficients. For example, the person-specific peak performance, a_i, parameters are assumed to follow a normal distribution, parametrized as:

$$a_i \sim N(\boldsymbol{x}_i \boldsymbol{\beta}_a, \sigma_a),$$

where $\boldsymbol{x}_i$ is a vector with a set of person-specific predictors such as sex (i.e., male or female) and MCI status, and with 1 as its first element (for the intercept). Vector $\boldsymbol{\beta}_a$ contains the corresponding regression coefficients. Specifically, the first coefficient of $\boldsymbol{\beta}_a$, that is $\beta_{a,\,int}$, takes the role of an intercept and quantifies the group (population)-level peak performance, while the rest of the regression coefficients correspond to the effects of the predictors in $\mathbf{x}_i$. In the analyses below we used age at baseline, MCI status, sex, and years of education as predictors. For example, regression coefficient $\beta_{a,\,age}$ quantifies the association between peak (asymptotic) performance and age at baseline, regression coefficient $\beta_{a,\,MCI}$ quantifies the association between peak performance and MCI status, regression coefficient $\beta_{\triangle,\,MCI}$ quantifies the association between change in peak performance and MCI status, and so on. Parameter σ_a quantifies residual variation in standard deviation units– that is individual differences remaining after the predictors are accounted for.

In the analyses below, similar specifications were made for $\triangle_i$, r_i, r_i^*, and $\sigma_{e,i}$, as well. The gain parameters were also made person-specific and assigned group-level distributions: $g_i \sim N(\mu_g, \sigma_g)$, where μ_g is the group-level mean gain across bursts, representing the average rate of gain in the sample throughout the study. The warm-up gain parameter g_i^* is assigned a group-level distribution that follows the same logic. However, these gain parameters were not regressed on person-level predictors the same way as the other parameters, as we did not expect them to be meaningfully related to our chosen set of predictors.

Finally, note that we are not specifying any particular correlation structure on the person-specific parameters (i.e., random effects); we are not constraining the correlation to be 0. All parameters, including regression coefficients, negative exponential model parameters and variances are estimated in a Bayesian framework, introduced next.

Casting the multilevel double exponential model in a Bayesian framework

In the Bayesian framework model parameters are thought of as random variables that have their own probability distributions. Bayesian modeling focuses on the estimation of posterior probability distributions (i.e., posteriors) based on available data (interpreted through a likelihood function) and prior probability distributions (i.e., priors) on the model parameters. Prior probability distributions are mathematical summaries of any already existing information on the model parameters. All inference is conditional on the priors, and they need to be specified before seeing the data to genuinely reflect the already existing information available on the parameters. The prior distributions for this study were chosen to be minimally informative, reflecting only the plausible, theoretical range of the parameters. This involved truncating distributions to match the parameter's range. For example, given that reaction times cannot be negative, peak performance of RT also cannot be negative, therefore its prior was truncated at 0 so that it cannot take on negative values.

Population means were given a prior that was distributed normally with 0 mean and standard deviation 10, truncated at 0 for across study (i.e., across bursts) and within-burst gain parameters, such as:

$$\mu_g \sim N(0,\ 10) \text{ and } \mu_{g*} \sim N(0,\ 10).$$

Regression coefficients (except for intra-individual standard deviation) were given the same priors, for example:

$$\beta_{a,age} \sim N(0, 10),$$

was specified for the coefficient linking asymptote (peak performance) and age. For the intra-individual standard deviation, we chose a somewhat tighter normal distribution with standard deviation equal to 1 to reflect that the likely range of this parameter was between 0 and 1, for example as:

$$\beta_{\sigma_e,age} \sim N\,(0,1)\,,$$

for the association between intra-individual standard deviation and age.

The group-level standard deviation parameters that reflect the heterogeneity across individuals were truncated to be on the positive real line and were specified as:

$$\sigma_r \sim N(0,\ 10),$$

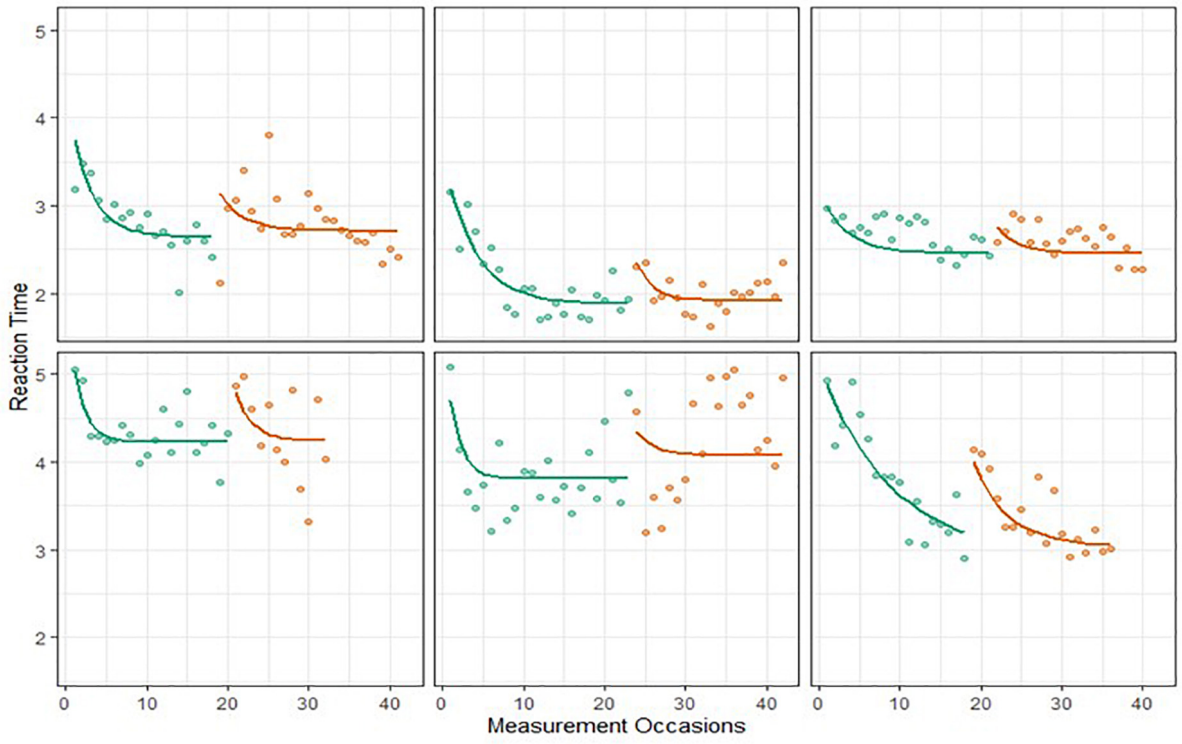


FIGURE 6
Six Einstein Aging Study (EAS) participants' symbol search data and predicted BDEM trajectories.

where σ_r could be substituted with σ_{r*}, σ_g, σ_{g*}, σ_a, and σ_δ. The standard deviation of the intra-individual standard deviation was specified as:

$$\sigma_{\sigma_e} \sim N(0, 1),$$

with range again truncated to the positive real line.

Implementation of double exponential model

The Bayesian double negative exponential model was implemented in Stan (Stan Development Team, 2022) called from R via rstan (Stan Development Team, 2020) – the Rscript for the estimation is available via OSF[1].

The results below were based on 60,000 samples from the posterior probability distributions of each model parameter. Specifically, we ran 6 parallel chains drawing 12,000 samples each, from which 2,000 per chain were discarded as warm-up iterations, resulting in 60,000 total iterations for each parameter. Sampling was performed via Markov chain Monte Carlo (MCMC) algorithms implemented in Stan. We checked MCMC sampling quality by calculating effective sample size (ESS) and $\hat{R}$ statistics. ESS quantifies the number of independent pieces of information in the posterior distribution and should be at least 100, but preferably around 1,000 to get reliable interval estimates. The $\hat{R}$ statistic is indicative of convergence of the sampled values, and values above 1.1 signal issues with convergence (Gelman et al., 2013). In our analyses, the ESSs for all parameters were above 100, and 98% of ESSs were also above 1,000 and all $\hat{R}$s were below 1.1.

Model fit

Symbol search task

We calculated the R^2 statistic to quantify the proportion of the variance in reaction times explained by the BDEM. For the current dataset the R^2 was 0.89, which supports a good fit of the model for the data. We also explored model fit visually by plotting model predicted trajectories over the data points, for every person, and concluded that the model followed the characteristics of the data satisfactorily. Specifically, we looked at whether the model predicted trajectory mimics the most important characteristics of the person-level data, which were whether (1) the height of the exponential curve overlaps with the first couple of observed data points, (2) the asymptote of the exponential curve overlaps with the best performances, and (3) the change in performance across the observations has an exponential shape. As an example, **Figure 6** shows 6 persons' raw data (dots) and model predicted trajectories, these were chosen

1 https://osf.io/h9yqk/?view_only=9e311fca177e462bbdb347c50736ae21

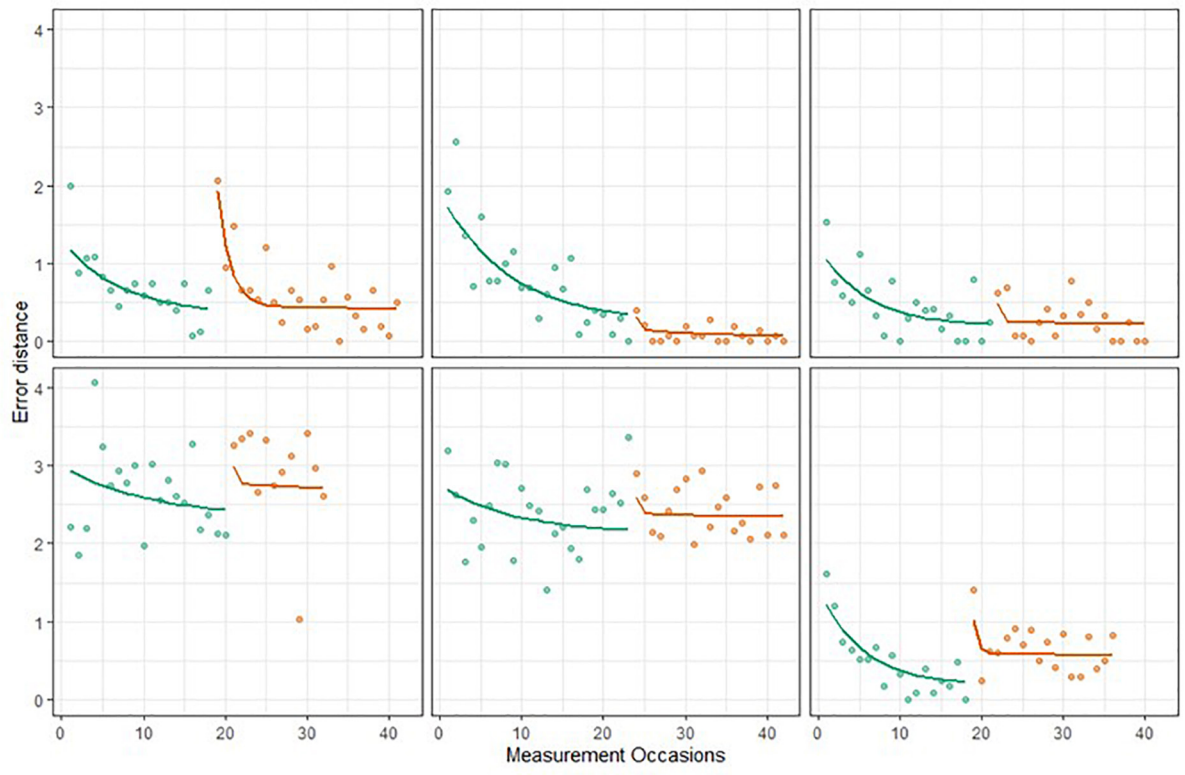


FIGURE 7
Six EAS participants' grid memory data and predicted Bayesian double exponential model (BDEM) trajectories.

to give representative illustration of the overall trends of the data.

Dot memory task

The BDEM showed sufficient fit to the grid memory data, with $R^2 = 0.76$, and predicted trajectories showing acceptable patterns; see Figure 7 for examples. However, we note that the fit of the BDEM was not as ideal for this data as for the symbol search data[2].

Results

We analyzed data for 318 participants, from which 171 completed both bursts, while the remaining participants had only Burst 1 data. The mean age of the sample at baseline was 77.45 (4.83) years and 67% were female (n = 104 male, and n = 214 female). The sample was racially and ethnically diverse with 45.9% (n = 146) identified as non-Hispanic Whites, 39.9% (n = 127) as non-Hispanic Blacks, 9.7% (n = 31) as Hispanic Whites, 2.8% (n = 9) as Hispanic Blacks, 1.3% (n = 4) as Asian, and 0.3% (n = 1) as more than one race/ethnicity. The mean education of the sample was 14.98 (3.55) years. On that basis of the neuropsychological assessment and criteria described above, 31.8% (n = 101) participants were classified as having MCI at baseline. There was no significant difference between those who completed both bursts and those with only Burst 1 in terms of age, years of education, race/ethnicity, or sex. But the group with only Burst 1 data was significantly more likely to be classified as MCI at Burst 1 (40.41% v. 24.42%, $p = 0.003$; please see table with all comparisons on the paper's OSF site). However, the BDEM mixed effects models we used can handle the missing data under the assumption that the data are missing at random (MAR). That is, the missing data process may depend on the predictors such as MCI status, covariates and the observed EMA cognitive outcomes at Burst 1. The only requirement for the missing data process is that conditional on MCI, covariates and Burst 1 EMA cognitive data, the missing data at Burst 2 must be independent of the unobserved Burst 2 cognitive performance.

The symbol search task

We note that we decided to scale the response times in seconds to keep the above specified prior settings in the

2 Additionally, plots showing the raw data averaged across persons and the corresponding model fits for both the symbol search task and dot memory task can be found on the above referenced OSF site. These plots showed satisfactory fit of the BDEM on the group level.

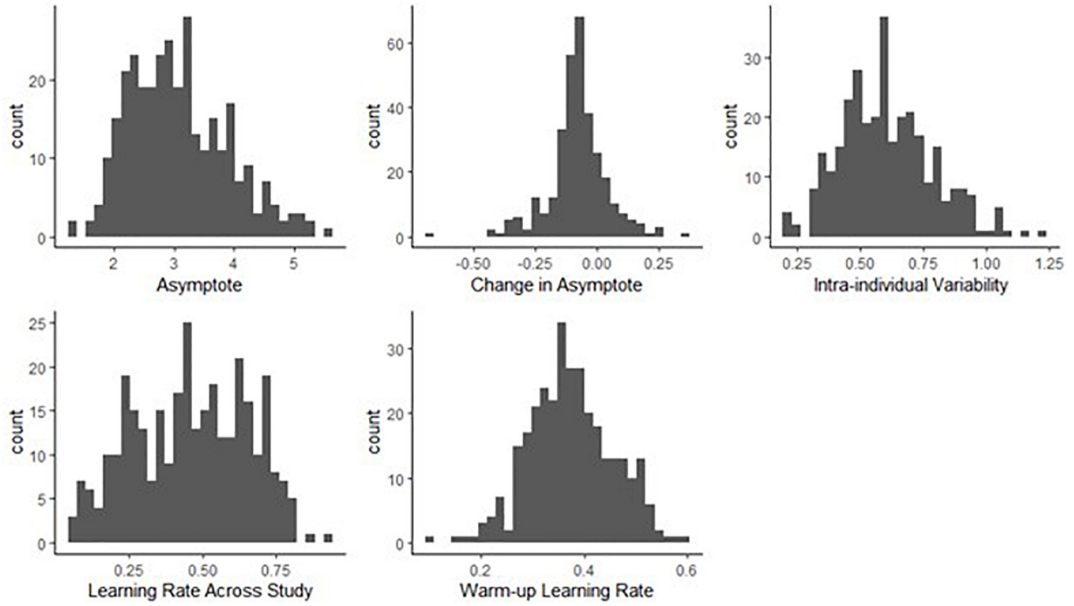


FIGURE 8
Histograms of person-specific estimates for key BDEM parameters based on data from the symbol search task. The horizontal axis shows the (binned) parameter values while the vertical axis displays the frequency of occurrence of that value among participants.

estimation algorithm the same for the symbol search and for the grid memory task data.

We did an initial data exploration by comparing the differences in Burst 1 and Burst 2 manifest performances (i.e., no BDEM). For every person, we calculated their average Burst 1 and Burst 2 reaction times, and created a difference score based on these (Burst 2 – Burst 1) to see how their performance changed across time. On average we found a 0.16s improvement in reaction times (M = −0.16, 95% CI: [−0.20, −0.12]), which was significantly different from 0 (t = −7.46, df = 170, p = 4.28e−12). This would suggest that on average participants got substantially faster in their reaction times in a year's time (between the two bursts) on this task. However, analysis based on these simple aggregates is confounded by practice effects. Next, we discuss how fitting the BDEM to this data showed different results.

Group-level (population) estimates and individual differences in the asymptote, change of asymptote, learning rates and intra-individual variability parameters based on the Bayesian double exponential model

We found a considerable amount of individual variation in asymptote, change of asymptote, learning rates and intra-individual variability parameters. Figure 8 shows the distributions of the person-specific point estimates of these parameters. Correspondingly, Table 1 shows their group-level averages (population mean estimates, e.g., $\beta_{a,int}$ for asymptote) and the amount of individual differences in them (heterogeneity in terms of population standard deviation estimates, e.g., σ_a, for asymptote). The column labeled "Mean" displays a point estimate for these parameters based on their posteriors, while the column labeled "PSD" shows the corresponding standard deviation around this point estimate, quantifying standard error.

TABLE 1 Group-level (population) estimates of Bayesian double exponential model (BDEM) parameters based on data from the symbol search task.

Process parameter	Mean	*PSD*
Asymptote averaged across individuals	2.83	0.06
Heterogeneity in asymptote (SD)	0.75	0.04
Change in asymptote averaged across individuals	−0.07	0.03
Heterogeneity in change in asymptote (SD)	0.19	0.02
Intra-individual variability averaged across individuals	0.56	0.02
Heterogeneity in intra-individual variability (SD)	0.18	0.01
Learning rate across study, averaged across individuals	0.49	0.04
Heterogeneity in learning rate across study (SD)	0.27	0.02
Warm-up learning rate averaged across individuals	0.39	0.05
Heterogeneity in warm-up learning rate (SD)	0.14	0.03

PSD indicates posterior standard deviation of the estimates, which quantifies standard error.

We can see that on average, the asymptote (i.e., peak performance, $\beta_{a,\ int}$) was 2.83 s (M = 2.83, PSD = 0.06) on this task. This intercept value (and the ones below) is the across person average corresponding to a participant who does not have the MCI status, who is male, and whose age and years of

education are at the sample mean level. There was considerable between-person variability in the asymptote, as shown by the standard deviation estimate (M = 0.75, PSD = 0.04) and the histogram of the person-specific asymptote estimates (first plot of Figure 8).

The (across-person) average difference in peak performance (asymptotes) between the first and the second bursts ($\beta_{\triangle,\ int}$) was −0.07 s (again, this corresponds to a participant who does not have the MCI status, who is male, and whose age and years of education are at the sample mean level), and it was credibly negative (M = −0.07, PSD = 0.03). This suggests that even when the retest effects were accounted for, there was an improvement in reaction time performance across bursts. However, the individual differences were considerable, as shown in the second plot of Figure 8: for example, for some participants, there was actually a slowing in reaction times, as shown by their positive change in peak performance estimate. In Section "Person-specific inference on the change in asymptotic performance via Bayesian probability distributions.," we will show how we can further scrutinize these individual-level estimates to get a probability estimate on whether the detected change represents credible decline in cognitive performance. Finally, we also note that 171 participants did not have second burst data yet, therefore their change estimates were informed by the population mean so they were all concentrated around −0.07.

The average intra-individual variation in RT ($\beta_{\sigma_e,\ int}$) was 0.56 s (M = 0.56, PSD = 0.02), with a large amount of variation across participants, quantified by the group-level standard deviation of the intra-individual variation parameter (M = 0.18, PSD = 0.01) and illustrated in the third plot of Figure 8. This suggests that individuals differ from each other considerably in terms of how much their cognitive performance fluctuates across the days.

Finally, with respect to the learning rate, a quick visual assessment of the plots in the second row of Figure 8 reveals that person-specific learning rates across study (between bursts) tend to be somewhat higher than the within-burst (warm-up) learning rates (see also corresponding entries in Table 1: M = 0.49 vs. M = 0.39); however, we can again see considerable amount of individual differences. Next, we look at the results of regressing these parameters on predictors to identify the sources of the individual differences.

Explaining sources of individual differences in the asymptote, change of asymptote, and intra-individual variability parameters with the Bayesian double exponential model

The person-specific asymptote, change in asymptote, intra-individual variability, learning rate across study and warm-up learning rate parameters were regressed on predictors quantifying age at baseline (standardized to mean of 0 and standard deviation 1), MCI status (coded as 0 and 1), sex (with 0 for female and 1 for male) and years of education (standardized similarly). Reported effects of age and education were all corresponding to 1 SD unit increase (4.83 years for baseline age, 3.55 years for years of education). Results on the regression coefficients quantifying their associations are summarized in Table 2. Just like in Table 1, the column labeled "Mean" displays a point estimate for these parameters, while the column labeled "PSD" shows the corresponding standard error. The last two columns show the probability that the regression coefficient is below and above 0, respectively, based on the posterior probability mass. For a credible effect we want to see at least 95% (0.95) probability of being either entirely below 0 or entirely above 0. However, we will also discuss if there was moderate evidence for effects, defined as at least 90% (0.9) probability of being either entirely below 0 or entirely above 0 (but not reaching the threshold of 0.95 for credible effect).

The first part of Table 2 shows that individual differences in asymptote (peak performance) were credibly linked to age, MCI status and years of education. With older age at baseline, peak performance reaction times showed credible slowing (0.11 s for each standard deviation of age). With positive MCI status there is also on average a 0.73 s slower peak performance reaction time. In contrast, with more years of education, peak performance reaction times tended to be faster [0.12 s faster per 1 SD (3.5) increase in years of education]. We did not find evidence for differences based on sex.

The second section of Table 2 summarizes the results with respect to changes in peak performance over time – that is between the two bursts in the study that were separated on average by a year. We only found trending support for association between age and change in peak performance: for each additional year older at baseline, participants tended to show a 0.03 s slowing of peak reaction times across the two bursts. For this effect to be credibly different from 0 there was 0.94 probability, which is slightly below our 0.95 threshold for credible effect. None of the other predictors showed credible links with this parameter.

The third section of Table 2 shows that differences in intra-individual variability in performance across time were credibly linked to MCI status and years of education. With positive MCI status there was a 0.15 s increase in the variability (in standard deviation units), while participants with 1 SD increases in years of education tended to show 0.04 s less variation.

The last two sections of Table 2 summarize the links between the learning rate parameters (across study and warm-up) and the selected predictors. We found only one credible link: participants with older age at baseline tended to show faster warm-up rate, meaning that they reached their peak performance faster in the second burst (0.05 s faster per one standard deviation on change in age).

TABLE 2 Summary of links between cognitive performance characteristics of the symbol search task and selected explanatory variables.

Process parameter	Predictor	Mean	*PSD*	<0	>0
Asymptote	Age	0.11*	0.05	0.01	0.99
	MCI status	0.73*	0.10	0.00	1.00
	Sex	0.03	0.10	0.39	0.61
	Years of education	−0.12*	0.05	1.00	0.00
Change in asymptote	Age	0.03ˆ	0.02	0.06	0.94
	MCI status	−0.03	0.05	0.74	0.26
	Sex	0.01	0.04	0.37	0.62
	Years of education	−0.02	0.02	0.77	0.23
Intra-individual variability	Age	0.01	0.01	0.25	0.75
	MCI status	0.15*	0.03	0.00	1.00
	Sex	0.01	0.02	0.34	0.66
	Years of education	−0.04*	0.01	1.00	0.00
Learning rate across study	Age	0.01	0.02	0.38	0.62
	MCI status	−0.06	0.05	0.88	0.12
	Sex	−0.04	0.05	0.82	0.18
	Years of education	−0.01	0.02	0.68	0.32
Warm-up learning rate	Age	0.05*	0.03	0.03	0.97
	MCI status	−0.06	0.05	0.88	0.12
	Sex	0.01	0.05	0.41	0.59
	Years of education	−0.03	0.03	0.89	0.11

Estimates with an * are meaningfully different from zero (at least 95% probability of being either entirely above or below 0). Estimates with a ˆ denote moderate evidence for an effect (at least 90% probability of being either entirely above or below 0). SD indicates posterior standard deviation of the estimates. Column "<0"/">0" displays the probability of the parameter being smaller/larger than 0.

The grid memory task

We did an initial data exploration for the grid memory task—much like we did for the symbol search task—by comparing differences in manifest performance between Burst 1 and Burst 2. We created person-specific difference scores between Burst 1 and Burst 2 averages based on the error distance measure. Across participants we found an improvement across bursts, specifically 0.21 units less error (M = −0.21, 95% CI: [−0.26, −0.15]), which was significantly different from 0 (t = −7.72, df = 170, p = 9.404e−13). This would suggest that participants' memory performance improved in a year's time (between the two bursts) on this task. However, as before, these simple aggregates are confounded by practice effects. We discuss results from the BDEM next.

Group-level estimates (population) estimates and individual differences in the asymptote, change of asymptote, learning rates, and intra-individual variability parameters based on the Bayesian double exponential model

Similar to the symbol search task, we found considerable amount of individual variation in asymptote, change of asymptote, learning rates and intra-individual variability parameters. Figure 9 shows the distributions of the person-specific point estimates of these parameters. Correspondingly, Table 3 shows their group-level averages and the amount of individual differences in them (following the same logic as in Table 1).

We can see that on average, the asymptotic, peak performance ($\beta_{a,\,int}$) was 1.85 units of error (M = 1.85, PSD = 0.09) on this task and that there was considerable between-person variance in peak performance, as shown by the standard deviation estimate (M = 0.69, PSD = 0.03) and the histogram of the person-specific asymptote estimates (first plot of Figure 9).

The (across-person) average difference in asymptotes (peak performance) between the first and the second bursts ($\beta_{\triangle,\,int}$) was 0.06 units of error (M = 0.06, PSD = 0.04). As opposed to credible improvement in peak reaction times on the symbol search task, this represents trending evidence for decline in performance over time. The individual differences in this peak performance change were also considerable, as shown in the second plot of Figure 9: while for most participants there was some level of decline in performance, there were also some whose performance improved across bursts.

The average intra-individual variation ($\beta_{\sigma_e,\,int}$) was around 1 grid unit (M = 1.06, PSD = 0.02), with a large amount of variation across individuals, quantified by the group-level standard deviation of the intra-individual variation parameter (M = 0.16, SD = 0.01) and illustrated in the third plot of Figure 9. This provided further evidence that participants differ from each other considerably in terms of how much their cognitive performance fluctuates across the days.

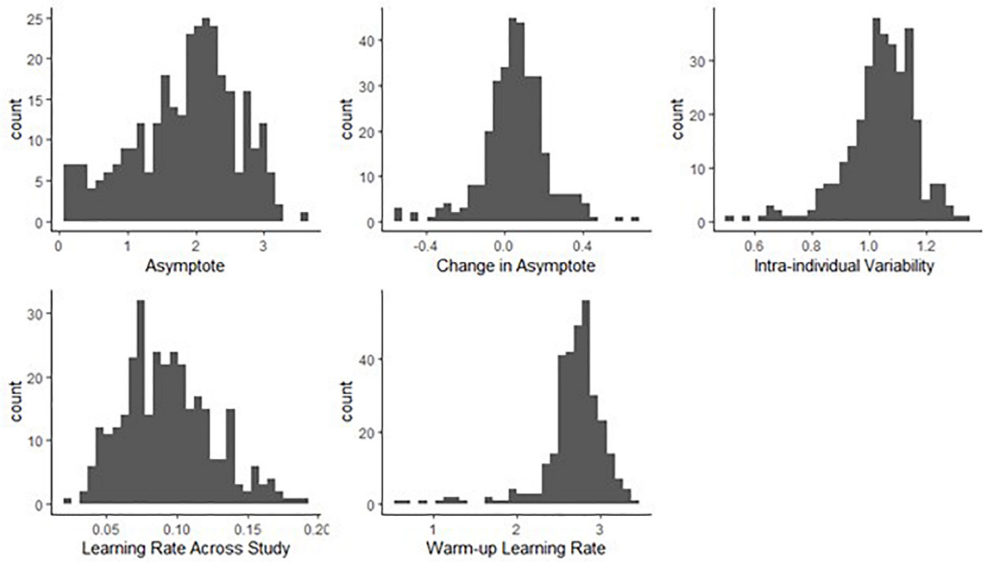


FIGURE 9
Histograms of person-specific estimates for key BDEM parameters based on data from the grid memory task. The horizontal axis shows the (binned) parameter values while the vertical axis displays the frequency of occurrence of that value among participants.

Finally, with respect to the learning rate, we found a different pattern than in the symbol search task: in this task the person-specific learning rates across study (between bursts) tended to be much lower than the within-burst (warm-up) learning rates, as illustrated in the second row of Figure 9 (see also corresponding entries in Table 3: $M = 0.09$ vs. $M = 2.67$); however, we can again see large individual differences. We again look at the results of regressing these parameters on predictors to identify the sources of the individual differences next.

Explaining sources of individual differences in the asymptote, change of asymptote, and intra-individual variability parameters with the Bayesian double exponential model

The first part of Table 4 shows that individual differences in asymptote (peak performance) were credibly linked to MCI status, sex and years of education. With positive MCI status the peak performance error rates were higher (on average by 0.44 units of error). In contrast, with being male and with more years of education, peak performance error rates tended to be lower (0.44 and 0.26 units of error, respectively). We did not find evidence for differences based on age.

The second section of Table 4 summarizes the results with respect to changes in peak performance over time – that is between the two bursts in the study that were separated on average by a year. We found credible support for association between age and change in peak performance: participants who were older at baseline tended to show worsening error rates (by 0.05 units of error) across the two bursts. In contrast, with more years of education participants tended to show improvement in error rate over time (0.04 units less). None of the other predictors showed credible links with this parameter.

TABLE 3 Group-level (population) estimates of BDEM parameters based on data from the grid memory task.

Process parameter	Mean	*PSD*
Asymptote averaged across individuals	1.85	0.09
Heterogeneity in asymptote (SD)	0.69	0.03
Change in asymptote averaged across individuals	0.06	0.04
Heterogeneity in change in asymptote (SD)	0.25	0.03
Intra-individual variability averaged across individuals	1.06	0.02
Heterogeneity in intra-individual variability (SD)	0.16	0.01
Learning rate across study, averaged across individuals	0.09	0.03
Heterogeneity in learning rate across study (SD)	0.04	0.01
Warm-up learning rate averaged across individuals	2.67	0.69
Heterogeneity in warm-up learning rate (SD)	0.99	0.23

PSD indicates posterior standard deviation of the estimates, which quantifies standard error.

The third section of Table 4 shows that differences in intra-individual variability in performance across time were credibly linked to age, sex, and education level: with older age, being male, and with more years of education, there was less variability (0.03, 0.08, and 0.02 in standard deviation units, respectively).

The last two sections of Table 4 summarize the links between the learning rate parameters (across study and warm-up) and the selected predictors. We found credible links only with across study learning rates, but the effect sizes were low. Older age at baseline, males, and participants with more years of education tended to be faster across study learning rates.

TABLE 4 Summary of links between cognitive performance characteristics of grid memory task and selected explanatory variables.

Process parameter	Predictor	Mean	*PSD*	<0	>0
Asymptote	Age	0.01	0.05	0.39	0.61
	MCI status	0.44*	0.10	0.00	1.00
	Sex	−0.44*	0.10	1.00	0.00
	Years of education	−0.26*	0.05	1.00	0.00
Change in asymptote	Age	0.05*	0.03	0.03	0.97
	MCI status	0.05	0.06	0.22	0.78
	Sex	−0.05	0.05	0.84	0.16
	Years of education	−0.04*	0.03	0.95	0.05
Intra-individual variability	Age	−0.03*	0.01	0.99	0.01
	MCI status	0.01	0.03	0.35	0.65
	Sex	−0.08*	0.03	1.00	0.00
	Years of education	−0.02*	0.01	0.96	0.04
Learning rate across study	Age	0.01*	0.10	0.03	0.97
	MCI status	−0.01	0.03	0.82	0.18
	Sex	0.03*	0.02	0.03	0.97
	Years of education	0.02*	0.01	0.01	0.99
Warm-up learning rate	Age	−0.08	0.14	0.73	0.27
	MCI status	0.03	0.86	0.56	0.44
	Sex	−0.01	0.33	0.52	0.48
	Years of education	−0.13	0.17	0.78	0.22

Estimates with an * are meaningfully different from zero (at least 95% probability of being either entirely above or below 0). Estimates with a ˆ denote moderate evidence for an effect (at least 90% probability of being either entirely above or below 0). SD indicates posterior standard deviation of the estimates. Column "<0"/">0" displays the probability of the parameter being smaller/larger than 0.

Person-specific inference on the change in asymptotic performance via Bayesian probability distributions

As stated before, the result of the Bayesian inference is a posterior probability distribution for every model parameter. Based on these distributions, probabilities on different ranges of the parameters can be calculated. This means, for example, decisions on the "significance" of regression effects do not need to be binary with an implausible null hypothesis of absolutely no difference. Instead, we can just make an informed decision by looking at the posterior probability distribution of the regression coefficient.

Inference can be done similarly for the person-specific parameters which are likely indicators of dementia risk, as on the change in peak performance across bursts. An example is shown in Figure 10 for symbol search and Figure 11 for grid memory data featuring the same six example participants as in Figures 6, 7. We can decide based on theoretical arguments whether a less than 0.01 s difference in peak performance (or 0.01 unit of error) represents a practically relevant effect. Using Monte Carlo integration, we can then calculate how much of the posterior mass falls above 0.01 (indicated with a vertical line in Figures 10, 11) – resulting in the probability of a practically relevant decline based on the participant's change in performance on a particular task between two bursts.

In Figure 10, we can see that for the participants in the first row and the first one in the second row, the posterior probabilities do not provide much evidence for practically relevant change – it is approximately the same amount of probability mass on both sides of 0. However, for the participant in the second plot of the second row of Figure 10, there is a 96% chance of such decline in symbol search performance, and the magnitude of decline is around 0.25 s, based on the peak of the posterior distribution (a more accurate point estimate can also be calculated). If we check the same participant's change in peak performance estimate from the grid memory task in Figure 11, there is a 94% probability of decline there, with the magnitude of decline being a bit less than 0.25 units of error, based on the peak of the posterior distribution. Inferences like this could be drawn for every person to evaluate their individual dementia risk.

As can be seen in Figure 11, the participants in the last row show high probabilities of cognitive decline based on their grid memory performance across the two bursts. For the participant in the third plot of the second row, there was already some support for decline on the symbol search task (70% chance, see Figure 10). Numerical probability estimates could also be combined together in a predictive modeling framework for efficient inference.

Discussion

Peak performance and changes in peak performance across bursts

In our analyses above, we aimed to isolate peak performance from retest effects in repeated measures of cognitive performance. We found that individual differences in the peak performance estimates were meaningfully related to the

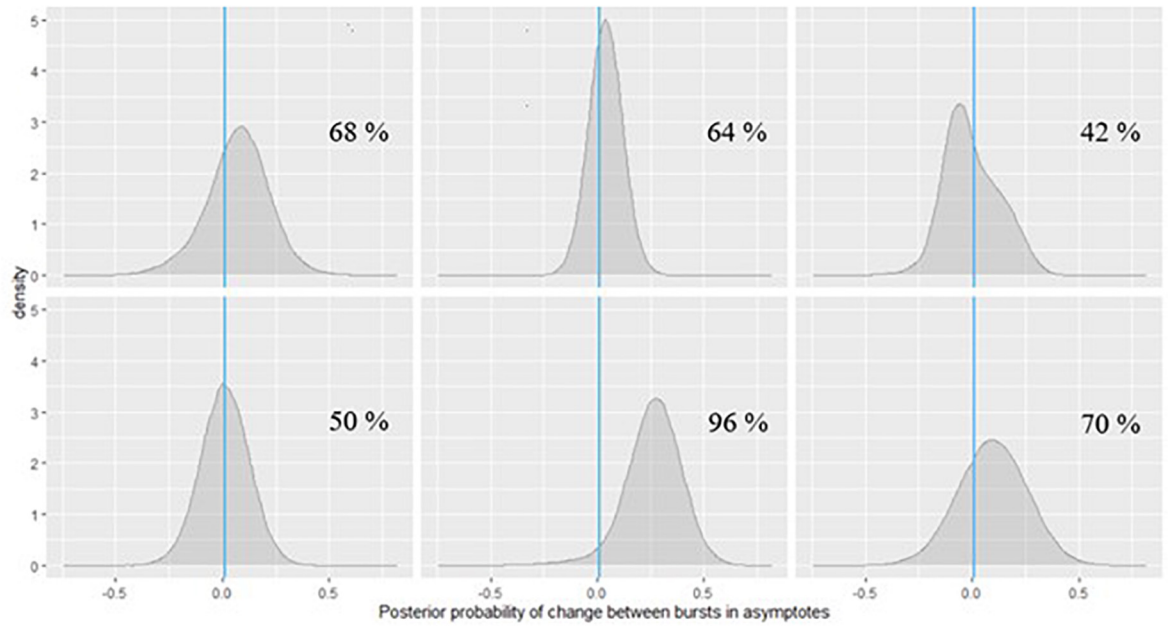


FIGURE 10
Posterior probabilities of change in the symbol search task performance for 6 EAS participants.

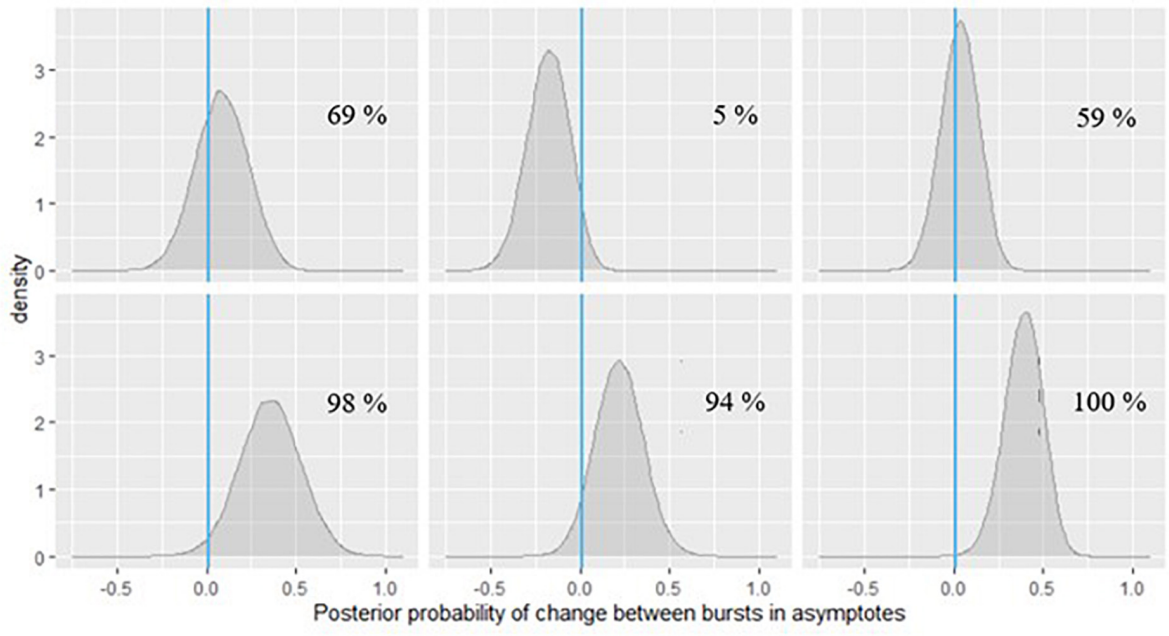


FIGURE 11
Posterior probabilities of change in the grid memory task performance for 6 EAS participants.

selected predictors. For example, MCI status was linked to decreased peak performance in both tasks.

When we explored the grid memory data by comparing burst averages, we found significant improvement across bursts. In contrast, the BDEM showed moderate evidence for decline in cognitive performance across bursts on this task. This suggests that disentangling learning processes from other latent cognitive changes is critical for this type of data. Individual differences in the change in peak performance across bursts were plausibly related to age (more error) and education (less error), further supporting the usefulness of our approach.

In contrast, on the symbol search task there was a 70-ms improvement across bursts in peak performance RT, even when retest learning effects were taken into account with the BDEM. However, this improvement is still smaller than the difference in burst averages (160 ms improvement), indicating that some retest learning was indeed accounted for by the BDEM. There are several possible reasons why we found improvement in peak RT on this task. It could be partly because we only have two bursts to examine change in peak performances across the years, so that we might not have had enough information to accurately capture the change process. Another reason for improvement in RTs on the symbol search task could be related to the fact that we were only modeling RTs from correct trials. Modeling all RTs in combination with accuracies for example in a drift diffusion model framework (see, e.g., Wagenmakers, 2009) could provide more insight.

Within-person variability in performance across days

While variability in performance is generally acknowledged in repeated assessments of cognitive performance, it is treated most often as a nuance. In the current study, we found that in the symbol search data, participants with MCI status showed more variability across days in their reaction times. Also, consistently across the two tasks, individuals with more years of education exhibited less variability. Paired with our previous findings that intra-individual variation in performance predicts MCI status, this may suggest that day-to-day variation reflects individual differences in cognitive reserve (Cerino et al., 2021).

Learning rates across study and within a burst

Learning effects confound the detection of cognitive change by biasing estimates of the underlying performance on a given assessment. In our study, we distilled these from peak performance estimated, but also considered them as potential indicators of cognitive change/decline given age- and disease-related impacts on brain subsystems that support learning. We extracted features of short- and long-timescale learning/retention in terms of within-burst or warm-up learning rate and across the study learning rate. On the symbol search task, we only found limited evidence (88% probability) of individuals with MCI status exhibiting slower learning; however, this effect was consistent for across study and warm-up learning rates (see **Table 2**). Surprisingly, on this task the only credible link was between learning and age, where participants who were older at baseline tended to show faster warm-up learning rate. Similar credible age effect was found in the grid memory data as well, although the effect was small and these participants also tended to have worse peak performance, therefore the steep learning might not indicate better brain health in this context.

Limitations and future directions

The double negative exponential model applied to measured burst data has the potential to provide a significant contribution toward accurately detecting and quantifying cognitive decline by disentangling practice effects from latent indicators of cognitive performance (i.e., asymptotic performance). It also provides clinically useful information in terms of personalized probabilities of impairment and decline for every individual, which can be useful to a clinician. We see several extensions of the BDEM approach for future projects. First, the BDEM parameter estimates on different tasks could be compared in terms of their predictive performance of neurodegenerative diseases. The goal is to optimize a model that has several of the key BDEM parameters as indicators, potentially from various cognitive domains (i.e., using more than one type of cognitive task). Second, while the current analysis did not yield promising results on linking learning process parameters with MCI status, it is possible that further exploration with indicators that are more specific to ADRD (such as blood biomarkers) could provide more insight. This is particularly relevant given that classification of MCI is a heterogenous classification, which as we highlighted in the introduction can have limited reliability. Finally, the BDEM could be combined with cognitive process models, such as the drift diffusion model that breaks down performances to meaningful cognitive characteristics. Combining such a drift diffusion modeling approach with the BDEM would allow us to simultaneously model and map learning features (e.g., learning rate) and changes in peak performance (and all associated random effects) onto cognitive (drift rate), and meta-cognitive (boundary separation) parameters.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Interested collaborators are asked to complete a concept proposal form (details for potential project, paper, or abstract) to be reviewed and forwarded to the Einstein Aging Study Steering Committee for consideration. Requests to access these datasets should be directed to MK, MPH at mindy.katz@einsteinmed.org. For additional information on data sharing requests for the Einstein Aging Study, see https://www.einsteinmed.edu/departments/neurology/clinical-research-program/eas/data-sharing.html.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board at Albert Einstein College of Medicine. The patients/participants provided their written informed consent to participate in this study.

Author contributions

ZO, JH, and MS conceptualized the theoretical model and research questions. ZO led the writing of the manuscript and conducted the data analysis. MS, MK, and CW designed the original study and supervised the data collection. All authors contributed to the writing of the manuscript, provided feedback on the data analysis, and made critical revisions of the manuscript for intellectual content.

Funding

ZO, MS, and JH was supported by the NIH (R56AG074208-01 and U2CAG060408). JH was supported by R00AG056670. MK and CW received funding from the NIH (NIA 2 P01 AG03949), the Leonard and Sylvia Marx Foundation, and the Czap Foundation. RZ was supported by the NIH (R03AG067006). KH was supported by the National Institute on Aging Grant T32 AG049676 to the Pennsylvania State University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baker, J. E., Lim, Y. Y., Pietrzak, R. H., Hassenstab, J., Snyder, P. J., Masters, C. L., et al. (2016). Cognitive impairment and decline in cognitively normal older adults with high Amyloid-β: A meta-analysis. *Alzheimer's Dement.* 6, 108–121. doi: 10.1016/j.dadm.2016.09.002

Broitman, A. W., Kahana, M. J., and Healey, M. K. (2020). Modeling retest effects in a longitudinal measurement burst study of memory. *Comput. Brain Behav.* 3, 200–207. doi: 10.1007/s42113-019-00047-w

Buschke, H. (1984). Cued recall in Amnesia. *J. Clin. Neuropsychol.* 6, 433–440. doi: 10.1080/01688638408401233

Cerino, E. S., Katz, M. J., Wang, C., Qin, J., Gao, Q., Hyun, J., et al. (2021). Variability in cognitive performance on mobile devices is sensitive to mild cognitive impairment: Results from the Einstein aging study. *Front. Digital Health* 3:758031. doi: 10.3389/fdgth.2021.758031

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., and Wagenmakers, E. J. (2009). A diffusion model decomposition of the practice effect. *Psychon. Bull. Rev.* 16, 1026–1036. doi: 10.3758/16.6.1026

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd Edn. London: Chapman and Hall/CRC, doi: 10.1201/b16018

Heathcote, A., Brown, S., and Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychon. Bull. Rev.* 7, 185–207. doi: 10.3758/BF03212979

Hoffman, L., Hofer, S. M., and Sliwinski, M. J. (2011). On the confounds among retest gains and age-cohort differences in the estimation of within-person change in longitudinal studies: A simulation study. *Psychol. Aging* 26, 778–791. doi: 10.1037/a0023910

Ivanova, I., Salmon, D. P., and Gollan, T. H. (2013). The multilingual naming test in Alzheimer's Disease: Clues to the origin of naming impairments. *J. Int. Neuropsychol. Soc.* 19, 272–283. doi: 10.1017/S1355617712001282

Jak, A. J., Bondi, M. W., Delano-Wood, L., Wierenga, C., Corey-Bloom, J., Salmon, D. P., et al. (2009). Quantification of five neuropsychological approaches to defining mild cognitive impairment. *Am. J. Geriatr. Psychiatry* 17, 368–375. doi: 10.1097/JGP.0b013e31819431d5

Jones, R. N. (2015). Practice and retest effects in longitudinal studies of cognitive functioning. *Alzheimer's Dement.* 1, 101–102. doi: 10.1016/j.dadm.2015.02.002

Katz, M. J., Wang, C., Nester, C. O., Derby, C. A., Zimmerman, M. E., Lipton, R. B., et al. (2021). T-MoCA: A valid phone screen for cognitive impairment in diverse community samples. *Alzheimer's Dement.* 13:e12144. doi: 10.1002/dad2.12144

Lawton, M. P., and Brody, E. M. (1969). Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living1. 9, 179–186. doi: 10.1093/geront/9.3_Part_1.179

Lövdén, M., Li, S. C., Shing, Y. L., and Lindenberger, U. (2007). Within-person trial-to-trial variability precedes and predicts cognitive decline in old and very old age: longitudinal data from the Berlin aging study. *Neuropsychologia* 45, 2827–2838. doi: 10.1016/j.neuropsychologia.2007.05.005

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., and Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of Dementia of the Alzheimer type. *Arch. Neurol.* 49, 1253–1258. doi: 10.1001/archneur.1992.00530360051017

Munoz, E., Sliwinski, M. J., Scott, S. B., and Hofer, S. (2015). Global perceived stress predicts cognitive change among older adults. *Psychol. Aging* 30, 487–499. doi: 10.1037/pag0000036

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *Int. Econ. Rev.* 25, 221–247. doi: 10.2307/2648877

Possin, K. L., Laluz, V. R., Alcantar, O. Z., Miller, B. L., and Kramer, J. H. (2011). Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. *Neuropsychologia* 49, 43–48. doi: 10.1016/j.neuropsychologia.2010.10.026

Rast, P., Macdonald, S. W., and Hofer, S. M. (2012). Intensive measurement designs for research on aging. *GeroPsych* 25, 45–55. doi: 10.1024/1662-9647/a000054

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Percept. Motor Skills* 8, 271–276. doi: 10.2466/pms.1958.8.3.271

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Soc. Pers. Psychol. Compass* 2, 245–261. doi: 10.1111/j.1751-9004.2007.00043.x

Sliwinski, M., Hoffman, L., and Hofer, S. (2010). "Modeling retest and aging effects in a measurement burst design," in *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development*, eds P. C. M. Molenaar and K. M. Newell (Washington, DC: American Psychological Association), 37–50.

Stan Development Team (2020). *RStan: the R interface to Stan. R package version 2.21.2*. Scarborough, ON: Stan Development Team.

Stan Development Team (2022). *Stan Modeling Language Users Guide and Reference Manual, VERSION*. Scarborough, ON: Stan Development Team.

Tombaugh, T. N., Kozak, J., and Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Arch. Clin. Neuropsychol.* 14, 167–177. doi: 10.1016/S0887-6177(97)00095-4

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *Eur. J. Cogn. Psychol.* 21, 641–671. doi: 10.1080/09541440802205067

Wechsler, D. (1987). *Instruction Manual for the Wechsler Memory Scale Revised*. New York, NY: Psychological Corporation.

Zhaoyang, R., Sliwinski, M. J., Martire, L. M., Katz, M. J., and Scott, S. B. (2021). Features of daily social interactions that discriminate between older adults with and without mild cognitive impairment. *J. Gerontol. Ser. B* 2021, gbab019. doi: 10.1093/geronb/gbab019

frontiers | Frontiers in Aging Neuroscience

TYPE Original Research
PUBLISHED 05 October 2022
DOI 10.3389/fnagi.2022.905329

OPEN ACCESS

EDITED BY
William Kremen,
University of California, San Diego,
United States

REVIEWED BY
Roos Jutten,
Massachusetts General Hospital and
Harvard Medical School, United States
Luis D. Medina,
University of Houston, United States

*CORRESPONDENCE
Ove Almkvist
ove.almkvist@ki.se

SPECIALTY SECTION
This article was submitted to
Alzheimer's Disease and Related
Dementias,
a section of the journal
Frontiers in Aging Neuroscience

RECEIVED 26 March 2022
ACCEPTED 15 August 2022
PUBLISHED 05 October 2022

CITATION
Almkvist O and Graff C (2022) Practice effects in cognitive assessments three years later in non-carriers but not in symptom-free mutation carriers of autosomal-dominant Alzheimer's disease: Exemplifying procedural learning and memory?
Front. Aging Neurosci. 14:905329.
doi: 10.3389/fnagi.2022.905329

Practice effects in cognitive assessments three years later in non-carriers but not in symptom-free mutation carriers of autosomal-dominant Alzheimer's disease: Exemplifying procedural learning and memory?

Ove Almkvist[1,2,3]* and Caroline Graff[2,4]

[1]Divisions of Clinical Geriatrics, Department of Neurobiology Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden, [2]Theme Inflammation and Aging, Karolinska University Hospital, Stockholm, Sweden, [3]Department of Psychology, Stockholm University, Stockholm, Sweden, [4]Divisions of Neurogeriatrics, Department of Neurobiology Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

Practice effects (PEs) defined as an improvement of performance in cognition due to repeated assessments between sessions are well known in unimpaired individuals, while less is known about impaired cognition and particularly in latent brain disease as autosomal-dominant Alzheimer's disease. The purpose was to evaluate the general (across tests/domains) and domain-specific PE calculated as the annual rate of change (ARC) in relation to years to the estimated disease onset (YECO) and in four groups of AD: asymptomatic mutation carriers (aAD, $n = 19$), prodromal, i.e., symptomatic mutation carriers, criteria for AD diagnosis not fulfilled (pAD, $n = 4$) and mutation carriers diagnosed with AD (dAD, $n = 6$) as well as mutation non-carriers from the AD families serving as a healthy comparison group (HC, $n = 35$). Cognition was assessed at baseline and follow-up about 3 years later by 12 tests covering six domains. The aAD and HC groups were comparable at baseline in demographic characteristics (age, gender, and education), when they were in their early forties, while the pAD and dAD groups were older and cognitively impaired. The results on mean ARC for the four groups were significantly different, small, positive, and age-insensitive in the HC group, while ARC was negative and declined with time/disease advancement in AD. The differences between HC and aAD groups in mean ARC and domain-specific ARC were not significant, indicating a subtle PE in aAD in the early preclinical stage of AD. In the symptomatic stages of AD, there was no PE probably due to cognitive disease-related progression. PEs were the largest in the verbal domain in both the HC and aAD groups, indicating a relationship with cognitive vulnerability.

The group-related difference in mean ARC was predominant in timekeeping tests. To conclude, the practice effect in over 3 years was suggested to be linked to procedural learning and memory.

KEYWORDS

practice effect, cognition, Alzheimer's disease, autosomal-dominant, normal ageing, progression

Introduction

The practice or retest or learning effect refers to a phenomenon that individuals, who are assessed a second time (not within the same session) with the same neuropsychological test(s), show improved performance in the absence of an intervention. The practice effect (PE) occurs both in normal individuals (Calamia et al., 2012; Machulda et al., 2013, 2017; Gross et al., 2015; Jutten et al., 2020; Samaroo et al., 2020; Lim et al., 2021) and in patients diagnosed with cognitive impairment (Machulda et al., 2013; Gross et al., 2018; Jutten et al., 2020). The occurrence of PE is so common that the absence of PE is considered a potential marker of disease progression (Zehnder et al., 2007; Hassenstab et al., 2015; Elman et al., 2018; Jutten et al., 2020; De Simone et al., 2021) and disease (Cooper et al., 2004; Zehnder et al., 2007). The common knowledge of PE is presented and summarized in large meta-analyses (Calamia et al., 2012; Duff and Hammers, 2020; Jutten et al., 2020).

There are a number of core issues regarding PE. The size has been estimated to be 0.2–0.6 standard deviations in normal individuals (Van der Elst et al., 2008) although smaller and larger estimates have been reported (Bartels et al., 2010; Scharfen et al., 2018a; Duff and Hammers, 2020). The size of the effect may vary with cognitive domain and the specific test (Calamia et al., 2012; Salthouse, 2015; Gross et al., 2018; Samaroo et al., 2020), premorbid/baseline level of cognitive function (Bartels et al., 2010; Arendasy and Sommer, 2017; Scharfen et al., 2019), test experience (Salthouse, 2015), task requirement (Arendasy and Sommer, 2017; Scharfen et al., 2018b), personality, e.g., anxiety (Jendryczko et al., 2019), length of retest intervals (Falleti et al., 2006; Calamia et al., 2012; Machulda et al., 2013; Salthouse, 2015; Scharfen et al., 2018b; Jutten et al., 2020), retest interval conditions, e.g., treatment (Jacobs et al., 2017; Jutten et al., 2020; Wang et al., 2020), demographic characteristics such as age (Salthouse, 2010; Calamia et al., 2012) and education (Bartels et al., 2010), type and severity of disease ranging from dementia (Cooper et al., 2004; Gross et al., 2015, 2018; Sánchez-Benavides et al., 2016), to mild cognitive impairment (Cooper et al., 2004; Bläsi et al., 2009; Calamia et al., 2012; Duff and Hammers, 2020), presence of comorbidity and risk factor for cognitive decline like APOE status and AD biomarkers (Zehnder et al., 2007; Machulda et al., 2013; Oltra-Cucarella et al., 2018; Jutten et al., 2020; Lim et al., 2021), and relationship with brain findings (Duff et al., 2017, 2018; Wilson et al., 2018; Jutten et al., 2020; Samaroo et al., 2020). Although there is a lot of knowledge regarding PE, there is still incomplete knowledge of serial assessments (Ivnik et al., 2000; Bartels et al., 2010; Heilbronner et al., 2010; Wilson et al., 2018; Scharfen et al., 2019; Jutten et al., 2020; Samaroo et al., 2020; Lim et al., 2021) and particularly on PE in asymptomatic latent disease in the preclinical stage of autosomal-dominant AD (adAD).

The purpose of the study was to investigate PE in repeated assessments of cognitive functions in carriers and non-carriers from six families with adAD. These individuals could be divided into four groups associated with varying degrees of present cognitive impairment: mutation carriers diagnosed with clinical dementia of AD (dAD), mutation carriers with symptoms but unfulfilled diagnostic criteria of AD, i.e., prodromal AD (pAD) and mutation carriers lacking symptoms, i.e., asymptomatic AD (aAD), who will develop Alzheimer's dementia in future, and finally non-carriers from adAD families serving as a healthy comparison group (HC). These individuals were followed with repeated clinical examinations including cognitive assessment of performance in five domains. These domains are selectively sensitive to brain involvement in AD; episodic memory is considered most sensitive and affected early in the disease course, while verbal knowledge is considered relatively stable and affected relatively late in the disease course.

In adAD, there is an option to characterize each individual in terms of disease advancement, i.e., years to estimated clinical onset (YECO; Bateman et al., 2012; Almkvist et al., 2017).

Following this outline, the first aim was to investigate the degree of PE measured as the annual rate of change (ARC) between two assessments in the four groups of AD participants (dAD, pAD, aAD, and HC). The hypothesis was that groups differed in relation to stage of disease progression showing PE in HC and possibly in aAD followed by the absence of PE in pAD and dAD. The second aim was to compare PE in specific cognitive domains/tests in HC and AD. The hypothesis was that PE varies between cognitive domains in relation to regional brain involvement linked to brain vulnerability in AD and aging. The third aim was to identify when PE is observed, or conversely when PE is not observed in disease progression in mutation carriers. The hypothesis was that PE is inversely associated with

disease progression (YECO) in mutation carriers and relatively unrelated to age in non-carriers (YECO).

Materials and methods

Participants

Adult members of six families carrying an early onset AD mutation were invited to a comprehensive clinical examination at the Memory Clinic, Karolinska University Hospital Huddinge, Sweden. Ninety-four individuals accepted to participate in the baseline examination and most individuals accepted follow-up examination ($n = 64$). There was no significant difference between the 94 and the 64 individuals in demographics (age, gender, and years of education), cognitive screening (MMSE), or mutation status (carrier/non-carrier) (all p-values of >0.1). The study concerned 29 mutation carriers from six adAD families and 35 non-carriers from the same six families.

Three families carried an APP mutation the Swedish *APP* K670N/M671L (Axelman et al., 1994), or the Arctic *APP* E693G mutation (Nilsberth et al., 2001), or the London *APP* V717I mutation (Goate et al., 1991). Three families carried a *PSEN1* I143T mutation (Keller et al., 2010); or the M146V mutation (Haltia et al., 1994); or the H163Y mutation (Axelman et al., 1998).

In autosomal-dominant AD families, it is possible to estimate each individual's time (years) to the expected clinical onset (YECO) of symptoms based on information from previous mutation carriers in each family. The family-specific mean age at onset of clinical symptoms is 36 ± 2 years for *PSEN1* I143T (Keller et al., 2010), 36 ± 3 years for *PSEN1* M146V (Haltia et al., 1994), 51 ± 7 years for *PSEN1* H163Y (Axelman et al., 1998; Thordardottir et al., 2015), 54 ± 5 years for APP_{SWE} (Axelman et al., 1994; Thordardottir et al., 2015), 56 ± 3 years for APP_{ARC} (Nilsberth et al., 2001; Thordardottir et al., 2015), and 57 ± 5 years for London *APP* V717I (Goate et al., 1991). For each participant, both mutation carriers and non-carriers, YECO was calculated as the difference between the individual's age at the time of the examination minus the family-specific age at clinical onset, i.e., YECO = the individual's present age—the expected family-specific onset of symptoms.

Procedure

All participants, mutation carriers and non-carriers, had a comprehensive clinical examination at each visit, which included somatic, neurological, psychiatric status, cognitive screening with the Mini-Mental Status Examination (MMSE; Folstein et al., 1975) and assessment of cognitive functions (see below), sampling of blood, urine and cerebrospinal fluid for standard analyses, and magnetic resonance imaging of brain anatomy. Although clinical examinations started as far back as 1993, essentially the same protocol was followed throughout the study.

Diagnosis

Based on the clinical examination at baseline, six mutation carriers were diagnosed as having dementia according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) (American Psychiatric Association, 1994) and AD according to the Alzheimer's Disease and Related Disorders Association (NINCDS-ARDRA) criteria (McKhann et al., 1984). These individuals constitute the dAD group. Mild Cognitive Impairment (MCI) was diagnosed following revised Petersen criteria (Winblad et al., 2004) and four mutation carriers were diagnosed as having MCI but criteria for AD were not fulfilled; they constitute the prodromal AD group. The 19 non-diagnosed mutation carriers had no AD-related symptoms and were cognitively unimpaired and considered to be asymptomatic although they were mutation carriers; they constitute the asymptomatic AD group.

At the first follow-up examination about 3 years after the baseline examination, 10 mutation carriers were diagnosed with AD (three pAD and one aAD at baseline developed dementia at follow-up), two mutation carriers were diagnosed as prodromal at follow-up, i.e., symptomatic, but AD criteria were not fulfilled (one aAD at baseline changed into pAD and one pAD remained as pAD). Seventeen mutation carriers were still evaluated as asymptomatic at follow-up. All individuals in the HC group were healthy and cognitively unimpaired. One healthy non-carrier had lifelong selective non-progressive cognitive difficulties due to a specific syndrome (topographical disorientation); the data for this participant were retained in the study but excluded for selectively impaired tests caused by the specific syndrome. Another non-carrier had been a boxer and participated in tournaments in young adulthood and later he had been affected by multiple small brain infarcts in middle age, which motivated to exclude him from the study.

Procedure

All individuals went through a standard comprehensive clinical examination, which included an interview with the participant and often with a close informant. The examination included somatic, neurological, and psychiatric statuses, sampling of blood, and cerebrospinal fluid [(CSF); (beta-amyloid, total, and phosphorylated tau)], brain imaging using magnetic resonance imaging (e.g., global

atrophy); and electroencephalography examination, and assessment of cognitive function (see below). The same protocol has been followed throughout the study during follow-up visits.

Assessment of cognitive function

Premorbid global cognitive function was assessed based on demographic information and reading test results (Tallberg et al., 2006). The following tests were used to assess cognitive domains: the Information and Similarities tests from the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981; Bartfai et al., 1994; WAIS-R) for verbal ability, the Block Design from WAIS-R and the Rey–Osterrieth Copy tests (Lezak et al., 2004) for visuospatial ability, the Digit Span from WAIS-R and the Corsi Span (Lezak et al., 2004) for short-term memory (STM), the Rey Auditory Verbal Learning test, including learning and retention after 30 min, and the Rey–Osterrieth retention after 30 min (Lezak et al., 2004) for verbal and visuospatial episodic memory, the Trail Making A test (Lezak et al., 2004) for attention and the Digit Symbol from WAIS-R and the Trail Making B (Lezak et al., 2004) for executive function. Raw scores were converted to *z*-scores using a reference group of healthy adults (Bergman et al., 2007). The z-scores are always directed so that positive values indicate a favorable performance.

Practice effect

The main outcome measure was the annual rate of change (ARC) defined as the unweighted score of the test result in z-score at the second visit—test result in z-score at the first visit divided by the time interval in years (one decimal) between the first and second visits for each of the 12 tests. Unweighted ARC score was computed for each domain; verbal (Information and Similarities, visuospatial (Block Design and Rey–Osterrieth Copy), STM (Digit Span and Corsi Span), episodic memory (RAVL learning and retention and Rey–Osterrieth retention), attention (TMTA), and finally executive (Digit Symbol and TMTB). Missing data occurred infrequently (total number of observations = 12 tests × 2 visits × 64 participants = 1,536, number of missing data = 92, 6.0%, half of the missing data occurred in RAVL retention due to inability, recorded as missing and not as 0).

The follow-up examination occurred after about 3 years (M±SD: 3.0 ± 3.5, range 0.6–20 years). Most participants had retest intervals between 2 and 4 years. The few extremely short and long retest intervals were due to participants' personal conditions.

Statistical analyses

Descriptive statistics were used for background characteristics. Bar graphs and scatter plots were used to visualize the results. A one-sample *t*-test was used to analyze if ARC deviated from 0. A one-way ANOVA was used to analyze group differences on ARC. A multivariate ANOVA was used to analyze the main effects of group and domain as well as the group-by-domain interaction on ARC.

Results

The background characteristics of participants in the four groups at the baseline visit are shown in Table 1. There was no significant difference between groups in age, gender, years of education, retest interval, premorbid IQ, and the number of *APOE* e4 alleles (all *p*-values of >0.1), while groups differed significantly in YECO ($F = 4.84$, df=3/59, $p < 0.01$, $\eta^2 = 0.20$) and global cognition assessed by MMSE ($F = 17.96$, df = 3/42, $p < 0.001$, $\eta^2 = 0.56$) in relation to the progression of AD.

The cognitive test results at baseline in each test for the HC and AD (aAD, pAD, and dAD) groups are shown in Supplementary Table 1. The groups differed significantly in 10 of the 12 tests and most strongly in episodic memory (RAVL learning, RAVL retention, and Rey– Osterrieth retention), executive function (Digit Symbol and TMTB), and visuospatial performance (Block Design) (see Table 2). The HC and aAD groups did not differ significantly in any test (all *p*-values of >0.1). The aAD and pAD groups differed significantly in two tests: TMTA ($t = 2.60$, df = 21, $p < 0.05$, Cohen's $d = 1.31$) and TMTB ($t = 3.50$, df = 19, $p < 0.01$, Cohen's $d = 1.94$). The pAD and dAD groups did not differ significantly in any test (all *p*-values of >0.1), although the mean *z*-scores were much poorer in the dAD group compared to the pAD group.

I. PE across cognitive tests in AD groups (aAD, pAD, and dAD) in comparison to HC

The practice effect was evaluated by the mean ARC in the 12 cognitive tests for the AD (aAD, pAD, and dAD) and HC groups. In Figure 1, a bar graph shows the mean ARC for the four groups. The hypothesis that the mean ARC equals 0 was rejected for the HC ($t = 2.89$, df = 34, $p < 0.01$, Cohen's $d = 0.49$) and dAD ($t = 4.57$, df = 4, $p < 0.01$, Cohen's $d = 2.04$) groups, but not for the aAD and pAD groups (*p*-value of >0.1). The mean ARC index differentiated the groups significantly ($F = 14.59$, df = 3/63, $p < 0.001$, $\eta^2 = 0.88$). The difference in mean ARC between the HC (M±SD: 0.05 ± 0.12) and aAD (M±SD: 0.01 ± 0.17) groups was not significant ($p > 0.1$), while the difference in mean ARC between the aAD (M±SD: 0.01 ± 0.17) and pAD (M±SD: −0.28 ± 0.44) groups was significant ($t = 2.37$, df = 22, $p < 0.05$, Cohen's $d = 1.19$). The difference in ARC

TABLE 1 Background characteristics at baseline in non-carriers (Healthy Comparison group, HC) and mutation carriers with AD (asymptomatic, prodromal and diagnosed AD).

	Non-carriers	Mutation carriers		
	HC	Asymptomatic	Prodromal	Diagnosed AD
N (females/males)	35 (17/18)	19 (6/13)	4 (1/3)	6 (2/4)
Age, y	39.7 ± 12.9	37.8 ± 10.1	51.3 ± 7.1	49.6 ± 7.1
Range, y	17–62	21–53	41–57	40–56
Education, y	11.0 ± 2.3	11.8 ± 2.1	12.5 ± 3.1	9.7 ± 1.8
Range, y	7–18	9–16	10–17	7–12
YECO at 1st visit, y	−9.5 ± 6.7	−12.8 ± 8.1	−0.1 ± 2.3	+0.6 ± 5.5
Range	−27 to +10	−26 to −3	−4 to +1	−6 to +6
Retest interval, y	3.4 ± 2.4	3.0 ± 1.9	3.3 ± 2.9	1.9 ± 0.8
Range, y	1–11	1–20	1–8	1–3
Premorbid IQ, iq-score	104 ± 7.7	110 ± 8.4	108 ± 9.8	111 ± 8.3
Range	91–116	94–123	97–116	97–111
MMSE, score	29.0 ± 1.6	28.8 ± 1.7	26.8 ± 1.5	21.0 ± 5.3
Range, score	23–30	27–30	24–28	14–26
APOE e4, frequency	10/35	7/19	0/4	2/6

TABLE 2 Practice effects expressed as the Annual Rate of Change (ARC) across cognitive domains at baseline in non-carriers (Healthy Comparisons group, HC) and mutation carriers varying in stage of AD disease course (asymptomatic AD and combined prodromal AD and dementia AD).

	Non-carriers	Mutation carriers			
Domain	HC	aAD	pAD and dAD	*P*	η^2
Mean cognition	+0.05 ± 0.11	+0.01 ± 0.17	−0.35 ± 0.33	***	0.40
Verbal	+0.19 ± 0.36	+0.10 ± 0.17	−0.13 ± 0.36	**	0.15
Visuospatial	+0.06 ± 0.34	−0.16 ± 0.68	−0.50 ± 0.49	*	0.11
STM	−0.02 ± 0.37	−0.02 ± 0.14	−0.26 ± 0.31	*	0.12
Episodic memory	+0.08 ± 0.26	−0.01 ± 0.16	−0.16 ± 0.28	Ns	0.08
Executive function	+0.04 ± 0.28	−0.05 ± 0.12	−0.41 ± 0.62	***	0.23
Attention	−0.02 ± 0.40	+0.04 ± 0.28	−0.37 ± 1.04	**	0.17

Significance and eta-square (η^2) from one-way (group) ANOVA on each domain.
ns, not significant; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

between the pAD (M±SD: −0.28 ± 0.44) and dAD (M±SD: −0.42 ± 0.21) was not significant ($p > 0.1$).

II. PE in cognitive tests/domains in HC and AD (aAD, pAD, and dAD) groups

The practice effect was evaluated by means of ARC in each cognitive test for the HC and AD groups; the descriptive data are shown in Supplementary Table 2. The four groups were significantly differentiated in 8 of the 12 tests. The practice effect was strongest in three tests, in which performance was measured by timekeeping (Digit Symbol, TMTA, and TMTB). The size of PE in the HC group varied between tests from the largest in the Similarities test ($z = +0.23$) followed by Information ($z = +0.15$) and RAVL learning and Rey–Osterrieth retention ($z = +0.11$) and Block Design ($z = 0.08$) and small in four tests (Digit Span, RAVL retention, Digit Symbol, and TMTB). Unexpectedly, the PE was negative in three tests (Rey–Osterrieth Copy, Corsi Span, and TMTA). The pairwise group differences were not significant in any test for the HC vs. aAD groups and the pAD vs. dAD groups (all *p*-values of >0.1) probably due to small sample sizes.

To increase the sample size in groups, the 12 test results were aggregated into six *a priori* cognitive domains: verbal (Information and Similarities), visuospatial (Block Design and Rey–Osterrieth Copy), STM (Digit Span and Corsi Span), episodic memory (RAVL learning, RAVL retention, and Rey–Osterrieth retention), executive function (Digit Symbol and TMTB), and attention (TMTA). The main outcome of a multivariate analysis (MANOVA) with domain as within

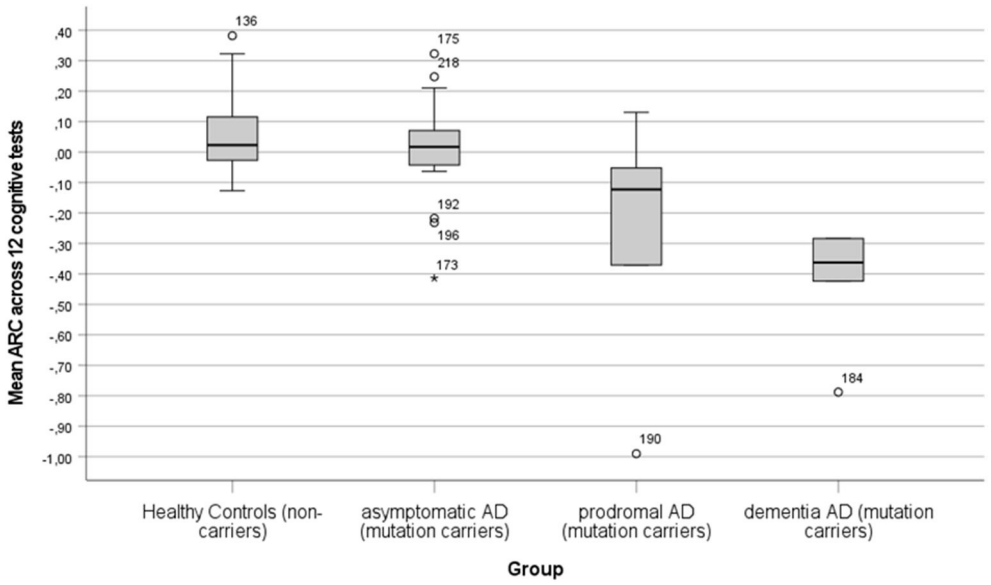


FIGURE 1
A bar graph showing the mean annual rate of change (ARC) with error bars in HC (non-carriers), aAD (asymptomatic mutation carriers), pAD (symptomatic mutation carriers, AD diagnosis nor fulfilled), and dAD (mutation carriers with AD diagnosis).

independent factor and group as between factor on ARC as dependent factor showed that the group effect was significant ($F = 7.14$, df = 3/55, $p < 0.001$, $\eta^2 = 0.28$), while the domain, as well as the group-by-domain interaction effects, were not significant (p-value of >0.1).

Still, the sample size was small in the pAD and dAD groups, so these groups were combined into a symptomatic AD (sAD) group encompassing mild and moderate cognitive impairment. The domain-specific ARC data for the three groups and the six cognitive domains are shown in Table 2. The group effect was significant in five of the six domains ($F = 10.89$, df = 2/56, $p < 0.001$, $\eta^2 = 0.28$). The domain effect was not significant ($p = 0.08$), and the group-by-domain interaction was not significant ($p > 0.1$). The addition of APOE e4 and/or education as covariates did not influence the outcome (p-value of >0.1).

The largest PE in the HC group was seen in the verbal domain ($z = +0.19$), and this was statistically different from 0 ($p < 0.01$). In the aAD group, PE was largest in the verbal domain ($z = +0.09$, $p < 0.05$). In the sAD group, some retest changes were negative and significant: visuospatial ($z = -0.39$, $p < 0.05$), STM ($z = -0.33$, $p < 0.05$), and executive ($z = -0.44$, $p<0.05$).

III. PE in relation to disease advancement in HC and AD (aAD, pAD, and dAD) groups

The relationship between PE and time of disease progression (YECO) was analyzed including all participants. It was hypothesized that PE is relatively stable in healthy individuals but varies with the degree of cognitive impairment and finally disappears in AD according to previous research. In Figure 2, a scatter plot is presented showing the mean ARC in relation to the time of disease advancement (YECO) for all participants divided into two groups, HC vs. AD. The graph visualized the regression line and the 95% confidence interval for the two groups. The regression for the HC group was linear and practically invariant in relation to time ($r = 0.02$). The equation for the HC group was ARC = 0.058 + 0.000 ^{x} YECO, i.e., PE = 0.058. The regression for the AD group (combining the aAD, pAD, and dAD into one AD group) was best described by a linear equation that was significant with YECO as a single predictor ($r = 0.53$, $F = 10.54$, df = 1/27, $p < 0.05$, $r^2 = 0.28$); the equation runs as follows: mean ARC = −0.267 – 0.530 × YECO. The intersection between the HC and AD groups occurred at YECO∼ −20, i.e., about 20 years before the estimated clinical onset. Looking at the intersection of confidence intervals, the HC and AD groups were separated at YECO∼ −12. Compared to the linear model, a curvilinear model was less powerful as well as models, in which other possible predictors (APOE e4 and/or years of education) were added. The alternative models did not increase the explanatory power.

Looking at Figure 2, a number of individuals both in the HC and AD groups were obvious outliers. In the HC group, three individuals had high positive ARC values (>0.30). In the AD group, there were at least three positive outliers (ARC > 0 and YECO > −4 close to the estimated onset) and five negative outliers far below the lower confidence line.

Next, the relationship was analyzed in each of the six domains. The non-linear regression of ARC in each domain on time (YECO) is reported as LOcally WEighted Scatterplot

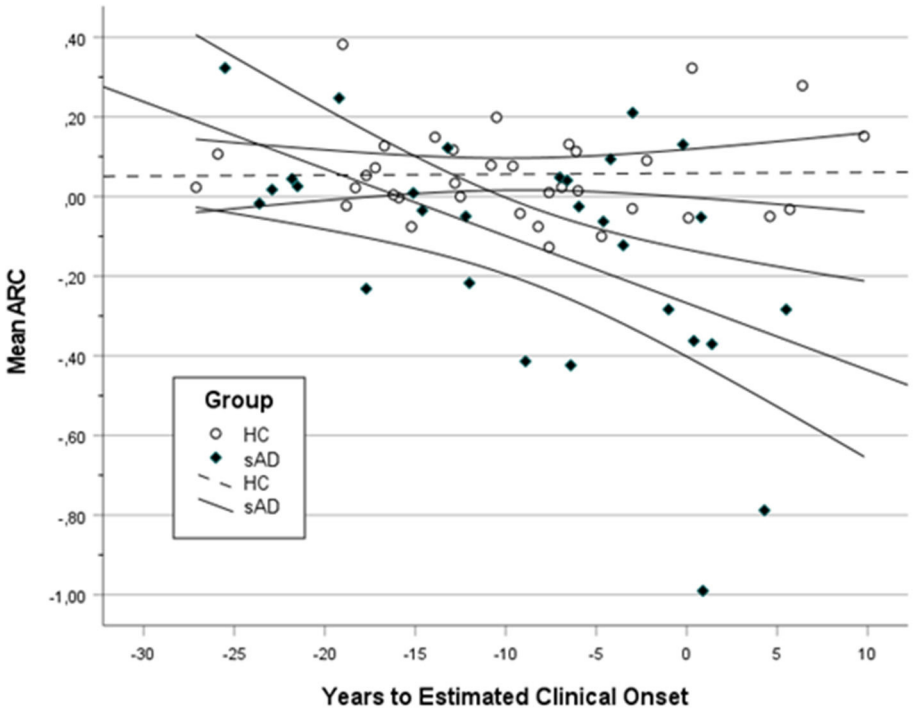


FIGURE 2
A scatter plot showing the mean annual rate of change (ARC) in HC and all AD (aAD, pAD, and dAD) in relation to years to estimated clinical onset (YECO) with a 95% confidence interval surrounding the linear regression line.

Smoothing lines, see Supplementary Figures 1–6. For the HC group, the regression lines were practically linear and parallel to the X-axis and ARC was very close to 0 in all domains, although relatively small for the entire time course that was covered by the study, see Figure 3 and Supplementary Table 1. For the AD group, the mean ARC was positive in the very early preclinical stage (YECO < -20), but later the mean ARC turned into negative ARC values in all six domains that increased with time, see Figure 3 and Supplementary Table 1. The decline started early in the executive and episodic memory domains about 10 years before clinical onset. The decline in other domains began later and was relatively close to the clinical onset of YECO.

Discussion

The study of PE with repeated cognitive assessments in mutation carriers and non-carriers from six families with autosomal-dominant Alzheimer's disease included mutation carriers varying in the stage of disease development in addition to healthy non-carriers. The carriers were diagnosed with Alzheimer's Disease (dAD), or prodromal AD expected to develop into dementia in the near future (pAD) or were lacking symptoms and regarded as asymptomatic although they will develop dementia in the distant future (aAD). All participants were examined at a memory clinic with a standardized protocol for patients with suspected dementia including a cognitive assessment with 12 tests covering six domains.

The first aim was to study PE measured as the annual rate of change (ARC) in cognition in the four groups (dAD, pAD, aAD, and HC). Results showed that PE aggregated across cognitive tests was positive in HC (M±SD: 0.056 ± 0.115), which is lower than reported in the previous literature (Van der Elst et al., 2008), probably depending on the length of the retest interval that was relatively long in this study (about 3 years in HC, aAD, and pAD, while it was about 1 year in dAD) compared short in many studies (Gross et al., 2018; Jutten et al., 2020; Samaroo et al., 2020). The hypothesis that mean ARC was equal to 0 was rejected in HC, but not in aAD implying that PE was absent or too small to be observed in aAD. The PE in HC was larger than in aAD individuals (M±SD: 0.007 ± 0.170), who lacked symptoms and were cognitively unimpaired despite carrying a mutation that will result in AD in the future. To speculate, the aAD individuals may have a subtle and unrecognized disturbance at this early stage about a decade prior to the estimated clinical onset. The results also showed that there was a negative PE in the dAD individuals, who were evaluated as mildly demented (MMSE M±SD: 21.0 ± 5.3) and the PE was lower than PE in the pAD group. This pattern of results supports that a practice effect exists in normal aging and is absent in clinically diagnosed AD

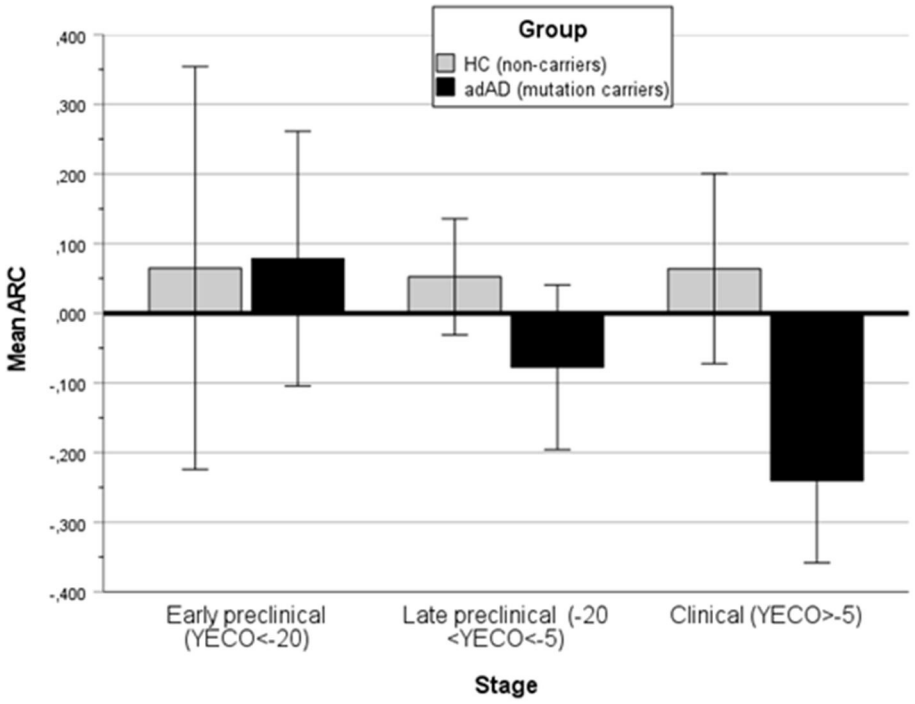


FIGURE 3
A bar graph showing the mean annual rate of change (ARC) in the HC and all AD groups (aAD, pAD, and dAD) in three stages of disease development: Early preclinical (YECO < −20), late preclinical (−20 < YECO < −5), and in the clinical stage around the estimated clinical onset.

as reported previously (Zehnder et al., 2007; Hassenstab et al., 2015; Elman et al., 2018; De Simone et al., 2021).

A few outliers in the mean ARC were observed. Two participants had extremely low mean ARC values (<-0.7, see Figure 1) and, in addition, they had short retest intervals that may have resulted in unreliable estimates that exaggerated the level of mean ARC. These mean ARC values are lower than the expected global cognitive decline (average across nine tests) previously estimated to be −0.43 in the mild stage of AD dementia (Almkvist and Bäckman, 1993). Finally, it should be pointed out that the negative ARC values represent values of annual progression of AD when practice effects are minor or absent.

The second aim was to study PE in specific cognitive tests with the expectation to find differences in correspondence with cognitive vulnerability associated with aging and disease (Cooper et al., 2004; Calamia et al., 2012; Salthouse, 2015). In order to improve stability across groups and tests, the pAD and dAD groups were combined into a symptomatic group and the 12 tests were aggregated into six domains (verbal, visuospatial, STM, executive, and attention). Now, the groups were differentiated in five of the six domains, and the effect of the domain was not significant, as well as the group-by-domain interaction. The largest power in differentiating the groups was obtained in the executive and attention domains that comprised timekeeping tests (Digit Symbol and TMTB as well as TMTA). This significant differentiation was obtained based on large negative and significant retest scores in the sAD group in executive and attention domains and not by positive PE in HC and/or aAD groups. In a similar vein, the preclinical decline in adAD in attention and executive function has recently been reported (Medina et al., 2021).

The significant and largest PE was observed in the verbal domain in the HC group in line with previous research (Calamia et al., 2012; Salthouse, 2015). PE was also positive in the verbal domain in the aAD group, although not significant. To speculate, the level of PE across cognitive domains in AD and HC is linked to cognitive vulnerability, i.e., lowest in the most vulnerable domains in AD considered to be episodic memory, executive, and visuospatial functions (Bateman et al., 2012; Almkvist et al., 2017). The largest PE was found in the verbal ability which is considered to be the least vulnerable domain in AD and in normal aging.

The third aim was to study the relationship between the size of PE and disease advancement estimated by YECO in the AD (mutation carriers with manifest and latent disease) and HC (healthy and cognitively unimpaired non-carriers). In the combined AD group, the relationship was linear and marked in the mean ARC. The change in mean ARC across time was about 0.06/year, which is less than the reported rate of change in

previous research (Van der Elst et al., 2008). The low mean ARC in this study could be due to the long retest interval compared to the shorter retest intervals used in previous research (Falleti et al., 2006; Calamia et al., 2012; Machulda et al., 2013; Salthouse, 2015; Scharfen et al., 2018a). The type of test (screening vs. domain-specific) may impose variation in PE (Gross et al., 2018).

In the AD group, the mean ARC began to deviate from the mean level in the HC group about 20 years prior to the clinical onset and the confidence interval for the AD and HC groups occurred when YECO was 10–15 years ahead of the estimated clinical onset. The intersection of regression lines and confidence interval in the HC and AD groups in this study on PE are in agreement with reports of trajectories in cognitive tests using separate measures in AD (Bateman et al., 2012; Almkvist et al., 2017; Medina et al., 2021). The finding that aAD individuals did not demonstrate a significant PE or a significant difference compared to HC individuals when assessed about 20 years ahead of the clinical onset is a novel finding.

It was observed that the level of PE varied a lot, particularly in the early preclinical stage of disease in the aAD sub-group of AD. However, the number of individuals in this group is too few to analyze this finding further. One possibility may be to analyze the relationship between mean ARC and a biomarker in general by including all cases, both non-carrier and carriers.

The main body of recent research on PE has focused on PE with short retest intervals and PE as a marker of cognitive progression, while relatively few studies have focused on PE observed at long retest intervals as in this study. It has been suggested that the mechanism of PE is related to various learning and memory processes, e.g., remembering test items, answers, and problems related to explicit declarative learning and retrieval processes related to the test content (Gross et al., 2018; McDermott, 2021). In contrast, the PE results of this study obtained with long test intervals and a comprehensive cognitive assessment are suggested to be related to procedural learning and memory when performing cognitive tasks repeatedly. A similar suggestion was proposed (named as a context effect) in a recent study of MMSE with a short test interval (Gross et al., 2018). In theory, this memory has been described as implicit and keeping knowledge relatively intact across time. The division of learning and memory into explicit declarative and implicit procedural systems varying in learning mode (consciously vs. unconsciously) and retrieval mode (recollection vs. acting) was suggested years ago (Squire, 2004; Squire and Dede, 2015). To this end, a meta-analysis has shown that performance in procedural learning and memory tasks appears to be preserved in individuals with aMCI and AD dementia compared to healthy older adults (De Wit et al., 2021). The distinction of performance in declarative and procedural memory in AD was supported in a large study on MMSE in patients with AD with reduced episodic memory by a PE at retest 4 months later (Gross et al., 2018). Recently, it was demonstrated that patients with MCI and cognitively unimpaired adults did not differ in performance of the classical procedural learning task (mirror tracking), while groups differed in typical episodic memory (the RAVL test) (De Wit et al., 2022).

In addition, a number of general factors operate during testing the second time and later, for instance relief from factors that hamper individuals from optimal cognitive performance (uneasiness, concerns of being tested) and factors that may improve performance the second time (coping/adaption associated with the experience of testing, change in strategies how to solve tasks) (Lievens et al., 2007). A favorable feature of the present study that was the complete examination was a 2-day long visit, the tests were the same, the psychologist was the same, and personal was the same to a large extent over the years. Taken together, it is suggested that part of PE in the present study can be understood as an example of procedural learning and memory that promote performance in cognitive testing when repeated. Interestingly, the brain structures involved in procedural learning and memory are different from the structures involved in AD (De Wit et al., 2021).

This study is based on a relatively small sample of mutation carriers and non-carriers from six adAD families; this is a disadvantage that has to be kept in mind. Particularly, the small sample size was obvious in the pAD and dAD groups. The material was analyzed both in terms of group comparisons and in terms of regression analysis to find converging results that could strengthen the conclusions. The fact that Alzheimer's disease was studied in four groups defined on genetics from no disease in HC to the asymptomatic stage, across mild and finally marked cognitive impairment in AD represents a favorable and unique feature of this study in contrast to other studies with clinically defined disease stages (Calamia et al., 2012; Duff and Hammers, 2020; Jutten et al., 2020). It is also a favorable feature that the retest interval was long and that cognition was studied extensively with several tests from six cognitive domains. This made it possible to compare PE across cognitive domains in interaction with stages of AD development and in relation to the estimated remaining time to the clinical onset of AD.

There are some implications of the present findings for clinical application and research. If the expected practice effects of repeated cognitive testing were not considered, previous results in follow-up clinical examinations and longitudinal studies may need to be reinterpreted. Furthermore, clinical trials may have come to incorrect conclusions on the effects of treatment if the PE phenomena were not regarded. However, the size of PE and the influence of covariates on PE has to be established in future research before it could be used in research and clinical application. The potential benefit of absent PE in short retest intervals as a marker of cognitive decline in aging and mild disease has been well documented in previous research

(Zehnder et al., 2007; Hassenstab et al., 2015; Elman et al., 2018; Jutten et al., 2020; Samaroo et al., 2020; De Simone et al., 2021). Finally, the mechanism of PE is not well understood. This fact makes it necessary to study both task-related cognitive factors as well as covert affective reactions.

To conclude, PE measured as ARC based on long retest intervals (about 3 years) were found in healthy and cognitively unimpaired middle-aged individuals (non-carriers from autosomal-dominant AD families) in age-insensitive cognitive domains. PE were also found in asymptomatic mutation carriers from AD families in the verbal cognitive domain when they were assessed long before the estimated clinical onset of AD. No PE, but a cognitive decline was obvious in symptomatic mutation carriers with mild cognitive impairment. In theory, PE are suggested to reflect that the person uses procedural learning and memory to master cognitive task demands in repeated testing.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

All participants were aware of their risk to develop AD. This information was given prior to the clinical examination. They also received genetic counseling in connection with the study and no one asked for information on their genetic status before the first visit or after the first visit. After the second follow-up examination, two asymptomatic participants opted for genetic testing after the completion of the examination. All subjects provided written informed consent to participate in the study. All examiners were blinded to the participants' mutation status. The study was approved by the Ethics Committee of Karolinska University Hospital at Huddinge and was conducted according to the Declaration of Helsinki and subsequent revisions.

Author contributions

Both authors fulfill the ICMJE criteria for authorship. Other collaborators have been mentioned and thanked for their assistance in parts of this study. Both authors have read and agreed to the final version of the manuscript.

Funding

This study was supported by grants from the Swedish Dementia foundation, Swedish Brain Foundation, Regional Agreement on Medical Training and Clinical Research (ALF) between Stockholm Region and Karolinska Institutet, Swedish Alzheimer Foundation, Stohnes Foundation, and Gamla Tjänarinnor Foundation.

Acknowledgments

The voluntary participation of all individuals has been gratefully acknowledged as well as the assistance of the research coordinators (Nathalie Asperén, Anne Kinhult-Ståhlbom, Catharina Roman), clinical examinations (Charlotte Johansson and Steinunn Thordardottir), APOE analyses (Jose Laffita-Mesa), and genetic guardian (Håkan Thonberg).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2022.905329/full#supplementary-material

References

Almkvist, O., and Bäckman, L. (1993). Progression in Alzheimer's disease: sequencing of neuropsychological decline. *Int. J. Geriatr. Psychiatr.* 8, 755–763. doi: 10.1002/gps.930080908

Almkvist, O., Rodriguez-Vieitez, E., Thordardottir, S., Amberla, K., Axelman, K., Basun, H., et al. (2017). Predicting cognitive decline across four decades in mutation carriers and non-carriers in autosomal-dominant Alzheimer's

disease. *J. Int. Neuropsychol. Soc.* 23, 195–203. doi: 10.1017/S1355617716001028

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders, 4th Edn.* Washington, DC: Author.

Arendasy, M. E., and Sommer, M. (2017). Reducing the effect size of the retest effect: examining different approaches. *J. Intell.* 62, 89–98. doi: 10.1016/j.intell.2017.03.003

Axelman, K., Basun, H., and Lannfelt, L. (1998). Wide range of disease onset in a family with Alzheimer disease and a His163Tyr mutation in the presenilin-1 gene. *Arch. Neurol.* 55, 698–702. doi: 10.1001/archneur.55.5.698

Axelman, K., Basun, H., Winblad, B., and Lannfelt, L. (1994). A large Swedish family with Alzheimer's disease with a codon 670/671 amyloid precursor protein mutation. *A clinical and genealogical investigation. Arch. Neurol.* 51, 1193–1197. doi: 10.1001/archneur.1994.00540240037013

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., and Ehrenreich, H. (2010). Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC Neurosci.* 11, 118. doi: 10.1186/1471-2202-11-118

Bartfai, A., Nyman, H., and Stegman, B. (1994). *Wechsler Adult Intelligence Scale revised: WAIS-R Manual.* Stockholm, Sweden: Psykologiförlaget.

Bateman, R., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N., et al. (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* 367, 795–804. doi: 10.1056/NEJMoa1202753

Bergman, I., Blomberg, M., and Almkvist, O. (2007). The importance of impaired physical health and age in normal cognitive aging. *Scand. J. Psychol.* 48, 115–125. doi: 10.1111/j.1467-9450.2007.00594.x

Bläsi, S., Zehnder, A. E., Berres, M., Taylor, K. I., Spiegel, R., and Monsch, A. U. (2009). Norms for change in episodic memory as a prerequisite for the diagnosis of mild cognitive impairment (MCI). *Neuropsychology* 23,189–200. doi: 10.1037/a0014079

Calamia, M., Markon, K., and Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin. Neuropsychol.* 26, 543–570. doi: 10.1080/13854046.2012.680913

Cooper, D. B., Lacritz, L. H., Weiner, M. F., Rosenberg, R. N., and Cullum, C. M. (2004). Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. *Alzheimer Dis. Assoc. Disord.* 18, 120–122. doi: 10.1097/01.wad.0000127442.15689.92

De Simone, M. S., Perri, R., Rodini, M., Fadda, L., De Tollis, M., Caltagirone, C., et al. (2021). A lack of practice effects on memory tasks predicts conversion to Alzheimer disease in patients with amnestic mild cognitive impairment. *J. Geriatr. Psychiatr. Neurol.* 34, 582–593. doi: 10.1177/0891988720944244

De Wit, L., Kessels, R. P. C., Kurasz,., A. M., Amofa, P., O'Shea, D., et al. (2022). Declarative learning, priming, and procedural learning performances comparing individuals with amnestic mild cognitive impairment, and cognitively unimpaired older adults. *J. Int. Neuropsychol. Soc.* 28:1–13. doi: 10.1017/S1355617722000029

De Wit, L., Marsiske, M., O'Shea, D., Kessels, R. P. C., Kurasz, A. M., DeFeis, B., et al. (2021). Procedural learning in individuals with amnestic mild cognitive impairment and alzheimer's dementia: a systematic review and meta-analysis. *Neuropsychol. Rev.* 31,103–114. doi: 10.1007/s11065-020-09449-1

Duff, K., Anderson, J. S., Mallik, A. K., Suhrie, K. R., Atkinson, T. J., Dalley, B. C. A., et al. (2018). Short-term repeat cognitive testing and its relationship to hippocampal volumes in older adults. *J. Clin. Neurosci.* 57, 121–125. doi: 10.1016/j.jocn.2018.08.015

Duff, K., and Hammers, D. B. (2020). Practice effects in mild cognitive impairment: a validation of Calamia. *Clin. Neuropsychol.* 27, 1–13. doi: 10.1080/13854046.2020.1781933

Duff, K., Hammers, D. B., Dalley, B., Suhrie, K. R., Atkinson, T. J., Rasmussen, K. M., et al. (2017). Short-term practice effects and amyloid deposition: providing information above and beyond baseline cognition. *J. Prev. Alzheimers Dis.* 4, 87–92. doi: 10.14283/jpad.2017.9

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., et al. (2018). Underdiagnosis of mild cognitive impairment: a consequence of ignoring practice effects. *Alzheimers Dement.* 4, 10, 372–381. doi: 10.1016/j.dadm.2018.04.003

Falleti, M. G., Maruff, P., Collie, A., and Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *J. Clin. Exp. Neuropsychol.* 28, 1095–112. doi: 10.1080/13803390500205718

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". *A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Goate, A., Chartier-Harlin, M. C., Mullan, M., Brown, J., Crawford, F., Fidani, L., et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 349, 704–706. doi: 10.1038/349704a0

Gross, A. L., Benitez, A., Shih, R., Bangen, K. J., Glymour, M. M., Sachs, B., et al. (2015). Predictors of retest effects in a longitudinal study of cognitive aging in a diverse community-based sample. *J. Int. Neuropsychol. Soc.* 21, 506–18. doi: 10.1017/S1355617715000508

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

Haltia, M., Viitanen, M., Sulkava, R., Ala-Hurula, V., Poyhonen, M., Goldfarb, L., et al. (1994). Chromosome 14-encoded Alzheimer's disease: genetic and clinicopathological description. *Ann. Neurol.* 36, 362–367. doi: 10.1002/ana.410360307

Hassenstab, J., Ruvolo, D., Jasielec, M., Xiong, C., Grant, E., and Morris, J. C. (2015). Absence of practice effects in preclinical Alzheimer's disease. *Neuropsychology* 29, 940–948. doi: 10.1037/neu0000208

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., and 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. *Clin. Neuropsychol.* 24, 1267–1278. doi: 10.1080/13854046.2010.526785

Ivnik, R. J., Smith, G. E., Petersen, R. C., Boeve, B. F., Kokmen, E., and Tangalos, E. G. (2000). Diagnostic accuracy of four approaches to interpreting neuropsychological test data. *Neuropsychology* 14, 163–177. doi: 10.1037/0894-4105.14.2.163

Jacobs, D. M., Ard, M. C., Salmon, D., P., Galasko, D. R., Bondi, M. W., et al. (2017). Potential implications of practice effects in Alzheimer's disease prevention trials. *Alzheimers Dement.* 3,531–535. doi: 10.1016/j.trci.2017.08.010

Jendryczko, D., Scharfen, J., and Holling, H. (2019). The impact of situational test anxiety on retest effects in cognitive ability testing: a structural equation modeling approach. *J. Intell.* 7, 22. doi: 10.3390/jintelligence7040022

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S. E., et al. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: a literature review. *Alzheimers Dement.* 12, e12055. doi: 10.1002/dad2.12055

Keller, L., Welander, H., Chiang, H. H., Tjernberg, L. O., Nennesmo, I., Wallin, A. K., et al. (2010). The PSEN1 I143T mutation in a Swedish family with Alzheimer's disease: clinical report and quantification of Aβ in different brain regions. *Eur. J. Hum. Gen.* 18, 1202–1208. doi: 10.1038/ejhg,.2010.107

Lezak, M. D., Howieson, D. B., and Loring, D. W. (2004). *Neuropsychological Assessment, 4th Edn.* New York, NY: Oxford University Press.

Lievens, F., Reeve, C. L., and Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *J. Appl. Psychol.* 92, 1672–1682. doi: 10.1037/0021-9010.92.6.1672

Lim, Y. Y., Baker, J. E., Mills, A., Bruns Jr, L., Fowler, C., Fripp, J., et al. (2021). Learning deficit in cognitively normal APOE ε4 carriers with LOW β-amyloid. *Alzheimers Dement.* 13, e12136. doi: 10.1002/dad2.12136

Machulda, M. M., Hagen, C. E., Wiste, H. J., Mielke, M. M., Knopman, D. S., Roberts, R. O., et al. (2017). Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. *Clin. Neuropsychol.* 31, 99–117. doi: 10.1080/13854046.2016.1241303

Machulda, M. M., Pankratz, V. S., Christianson, T. J., Ivnik, R. J., Mielke, M. M., Roberts, R. O., et al. (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo clinic study of Aging. *Clin. Neuropsychol.* 27, 1247–1264. doi: 10.1080/13854046.2013.836567

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Ann. Rev. Psychol.* 72, 609–633. doi: 10.1146/annurev-psych-010419-051019

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology* 34, 939–944. doi: 10.1212/WNL.34.7.939

Medina, L. D., Woo, E., Rodriguez-Agudelo, Y., Caparro Maldonado, H., Yi, D., Coppola, G., et al. (2021). Reaction time and response inhibition in autosomal dominant Alzheimer's disease. *Brain Cognit.* 147, 105656. doi: 10.1016/j.bandc.2020.105656

Nilsberth, C., Westlind-Danielsson, A., Eckman, C. B., Condron, M. M., Axelman, K., Forsell,. C., et al. (2001). The 'Arctic' APP mutation (E693G) causes

Alzheimer's disease by enhanced Abeta protofibril formation. *Nat. Neurosci.* 4, 887–893. doi: 10.1038/nn0901-887

Oltra-Cucarella, J., Sánchez-SanSegundo, M., and Ferrer-Cascales, R. (2018). Alzheimer's disease neuroimaging, initiative. Cognition or genetics? Predicting Alzheimer's disease with practice effects, APOE genotype, and brain metabolism. *Neurobiol. Aging* 71, 234–240. doi: 10.1016/j.neurobiolaging.2018.08.004

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology* 24, 563–572. doi: 10.1037/a0019026

Salthouse, T. A. (2015). Test experience effects in longitudinal comparisons of adult cognitive functioning. *Dev. Psychol.* 51, 1262–1270. doi: 10.1037/dev0000030

Samaroo, A., Amariglio, R. E., Burnham, S., Sparks, P., Properzi, M., Schultz, A. P., et al. (2020). Diminished learning over repeated exposures (LORE) in preclinical Alzheimer's disease. *Alzheimers Dement.* 12, e12132. doi: 10.1002/dad2.12132

Sánchez-Benavides, G., Peña-Casanova, J., Casals-Coll, M., Gramunt, N., Manero, R. M., Puig-Pijoan, A., et al. (2016). One-year reference norms of cognitive change in spanish old adults: data from the NEURONORMA sample. *Arch. Clin. Neuropsychol.* 31, 378–388. doi: 10.1093/arclin/acw018

Scharfen, J., Blum, D., and Holling, H. (2018a). Response time reduction due to retesting in mental speed tests: a meta-analysis. *J. Intell.* 6, 6. doi: 10.3390/jintelligence6010006

Scharfen, J., Jansen, K., and Holling, H. (2018b). Retest effects in working memory capacity tests: a meta-analysis. *Psychon. Bull. Rev.* 25, 2175–2199. doi: 10.3758/s13423-018-1461-6

Scharfen, J., Peters, J. M., and Holling, H. (2019). Retest effects in cognitive ability tests: a meta-analysis. *Intelligence* 67, 44–66. doi: 10.1016/j.intell.2018.01.003

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiol. Learn. Mem.* 82, 171–177. doi: 10.1016/j.nlm.2004.06.005

Squire, L. R., and Dede, A. J. (2015). Conscious and unconscious memory systems. *Cold Spring Harb. Perspect Biol.* 2, a021667. doi: 10.1101/cshperspect.a021667

Tallberg, I. M., Wenneborg, K., and Almkvist, O. (2006). Reading words with irregular decoding rules: a test of premorbid cognitive function? *Scand. J. Psychol.* 47, 531–539. doi: 10.1111/j.1467-9450.2006.00547.x

Thordardottir, S., Stahlbom, A. K., Ferreira, D., Almkvist, O., Westman, E., Zetterberg, H., et al. (2015). Preclinical cerebrospinal fluid and volumetric magnetic resonance imaging biomarkers in Swedish familial Alzheimer's disease. *J. Alzheimers Dis.* 43, 1393–1402. doi: 10.3233/JAD-140339

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., and Jolles, J. (2008). A large-scale cross-sectional and longitudinal study into the ecological validity of neuropsychological test measures in neurologically intact people. *Arch. Clin. Neuropsychol.* 23, 787–800. doi: 10.1016/j.acn.2008.09.002

Wang, G., Kennedy, R. E., Goldberg, T. E., Fowler, M. E., Cutter, G. R., and 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, e0228064. doi: 10.1371/journal.pone.0228064

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale Revised: WAIS-R Manual.* New York, NY: Psychological Corporation.

Wilson, R. S., Capuano, A. W., Yu, L., Yang, J., Kim, N., Leurgans, S. E., et al. (2018). Neurodegenerative disease and cognitive retest learning. *Neurobiol. Aging* 66, 122–130. doi: 10.1016/j.neurobiolaging.2018.02.016

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., et al. (2004). Mild cognitive impairment: beyond controversies, towards a consensus. *J. Int. Med.* 256, 240–246. doi: 10.1111/j.1365-2796.2004.01380.x

Zehnder, A. E., Bläsi, S., Berres, M., Spiegel, R., and Monsch, A. U. (2007). Lack of practice effects on neuropsychological tests as early cognitive markers of Alzheimer disease? *Am. J. Alzheimers Dis. Other Dement.* 22, 416–426. doi: 10.1177/1533317507302448

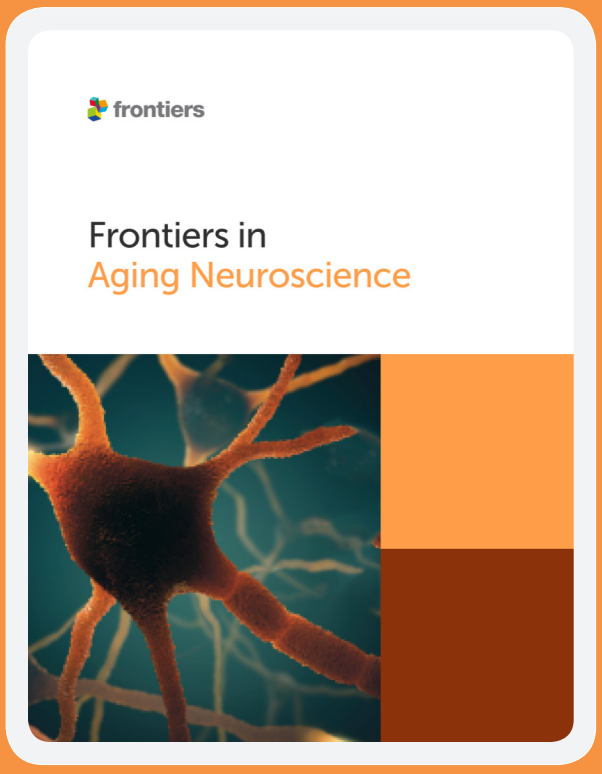
Frontiers in
Aging Neuroscience

Explores the mechanisms of central nervous system aging and age-related neural disease

The third most-cited journal in the field of geriatrics and gerontology, with a focus on understanding the mechanistic processes associated with central nervous system aging.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact