



Published in final edited form as:

J Clin Exp Neuropsychol. 2024 September ; 46(7): 630–643. doi:10.1080/13803395.2024.2400107.

Evaluating Practice Effects Across Learning Trials – Ceiling Effects, or Something More?

Dustin B. Hammers¹, Shreya Bothra¹, Angelina Polsinelli¹, Liana G. Apostolova¹, Kevin Duff²

¹Indiana University School of Medicine, Department of Neurology, Indianapolis, IN, USA

²Oregon Health and Science University, Department of Neurology, Portland, OR, USA

Abstract

Background: Practice effects (PE) are traditionally considered improvements in performance observed resulting from repeated exposure to test materials across multiple testing sessions. While PE are commonly observed for memory tests, this effect has only been considered in summary total scores. The current objective was to consider PE in summary total scores, individual learning trials, and learning slopes.

Method: One-week PE for individual trial and learning slope performance was examined on the BVMT-R and HVLIT-R in 151 cognitively intact participants and 131 participants with Mild Cognitive Impairment (MCI) aged 65 years and older.

Results: One-week PE were observed across all trials and summary total scores for both memory measures and diagnostic classifications, despite the potential for ceiling effects to limit improvement on retesting. PE were largest on the first trial relative to subsequent learning trials. This effect was diminished – but not eliminated – in participants with MCI. Conversely, no PE were observed for learning slope scores, which was counter to expectations and likely confounded by ceiling effects.

Conclusions: PE were present across learning trials but not learning slopes, and the initial learning trial at follow-up tended to benefit most from PE relative to subsequent learning trials. Ceiling effects appeared to influence PE for learning slopes more than learning trials. These results highlight the potential diagnostic utility of PE across individual learning trials and inform how they are distributed at follow-up, while also suggesting that learning slopes may be generally stable during longitudinal assessment.

Keywords

Reliable Change; Assessment; Neuropsychology; Practice Effects; Memory

Address correspondences to Dustin B. Hammers, PhD, Department of Neurology, Indiana University School of Medicine, Department of Neurology, 355 West 16th Street (GH4027), Indianapolis, IN, 46202. HammersD@iu.edu.

Conflicts: No authors associated with this project have reported conflicts of interest that would impact these results.

Consent: All authors have read and provided consent to be associated with this manuscript.

INTRODUCTION

Repeat cognitive assessment is a standard practice in clinical neuropsychology settings (Chelune & Duff, 2012), and is valuable when considering cognitive trajectory over time. For instance, patients with underlying etiology for Alzheimer's disease (AD) or another neurodegenerative condition would be expected to demonstrate consistent cognitive declines over time (Cohen, Reisberg, & Yaffee, 2022). In this context, serial assessment could provide diagnostic clarity in complex clinical presentations. Additionally, repeat assessment has an important role in evaluating the impact of interventions on cognition. The FDA approval of disease modifying treatments for AD (e.g., lecanemab) increases the need for post-treatment assessment of change over time. Additionally, treatment of common reversible causes of cognitive decline (e.g., cardiovascular disease, obstructive sleep apnea) may lead to cognitive improvements (Bubu et al., 2020; Gagnon et al., 2022). Further, conditions like stroke often require cognitive rehabilitation strategies that may directly improve cognition following their initiation (das Nair, Cogger, Worthington, & Lincoln, 2016).

Confidence in the ability to interpret change over time is paramount, as described in the situations above. Unlike most other medical procedures, exposing patients to testing materials at a baseline cognitive evaluation can improve cognitive test scores at a later evaluation (Holm, Wolfer, Pointeau, Lipsmeier, & Lindemann, 2022), whether or not those changes are associated with enhanced cognitive abilities. As a result, these "practice effects" (PE) can be considered a form of measurement error that poses a potential challenge to clinical interpretation of serial assessment results if not adequately considered (Hinton-Bayre, 2016). In contrast, evidence suggests that incorporation of PE as a clinical marker (i.e., "Do patients benefit from practice?") may provide clinicians with additional process-related tools to consider treatment response (Duff, Beglinger, Moser, Schultz, & Paulsen, 2010) and disease pathology/etiology (Duff et al., 2018). Altogether, PE are critical process scores for clinical neuropsychologists to account for during repeat assessment.

Benefit from practice may arise from the repeated assessment of any cognitive domain, though episodic memory measures tend to result in the highest PE (Calamia, Markon, & Tranel, 2012; Duff & Hammers, 2022) relative to other domains. PE from memory tests are also related to a variety of AD biomarkers, including hippocampal atrophy (Duff et al., 2018), hypoactivation using fluoro-2-deoxyglucose positron emission tomography (PET; Duff, Horn, Foster, & Hoffman, 2015), and β -amyloid deposition (Duff, Suhrie, Dalley, Anderson, & Hoffman, 2019). However, up to this point, investigation of memory PE has focused on summary scores for memory tests (e.g., the total items learned across tasks, or the total items retained upon a delay), with limited consideration of improvements due to practice during individual learning trials. As a result, it is currently unclear how improvements from one testing session to the next are distributed across the individual trials of a memory task. It is possible that *between*-session improvements (i.e., PE) result in incremental benefit across trials, or alternatively they may manifest predominantly during a particular trial during the follow-up session (e.g., the first trial on follow-up). Conversely, no PE may be present, driven in part by the possible impact of ceiling effects on repeated performance. Specifically, ceiling effects occur during memory testing when a participant

achieves the highest possible score during learning trials and cannot subsequently display improvements on later trials or test administrations. This has the highest potential to be observed in cognitively intact individuals displaying strong performances, with the result being that improvements (or PE) could be artificially attenuated because there is no more information for them to learn. For example, if a participant learns 12 of 12 items on a word list in their first attempt, then there is no opportunity for them to learn additional items during later exposures. As such, no PE would be present, though this would be due to psychometric constraints, not because the individual was unable to benefit from repeated exposure. Although this ceiling effect is a possibility, it is unknown whether ceiling effects would actually compromise the ability to understand PE across learning trials.

Relatedly, process scores are another way of evaluating learning above and beyond total summary scores. Research into learning slopes during multi-trial learning tasks has indicated that the rate of learning after the first trial can inform diagnostic decisions and treatment recommendations. For example, more shallow learning curves have been observed in a variety of clinical conditions, including those ranging from children with traumatic brain injury (Warschawsky, Kay, Chi, & Donders, 2005) to both middle aged (Hammers, Nemes, et al., 2023) and older adults with dementia due to AD (Gifford et al., 2015; Hammers, Suhrie, Dixon, Gradwohl, Duff, et al., 2022). Deficits in learning slope have also been associated with AD biomarkers, including increased β -amyloid burden (Hammers, Kostadinova, et al., 2023; Hammers, Pentchev, Kim, Spencer, & Apostolova, 2023) and tau deposition (Hammers, Lin, et al., 2023), and decreased hippocampal volume (Kristine B. Walhovd et al., 2020). As detailed in the Methods, a common learning slope metric is the Learning Ratio (LR), which is defined as the amount of information acquired after the first trial divided by the amount of information left-to-learn after trial one (Spencer, Gradwohl, Williams, Kordovski, & Hammers, 2020). As an example of the calculation of this LR metric, if a participant learned 2 words on a 3-trial 12-item word-list learning test at Trial One, 4 words at Trial Two, and 6 words at Trial Three, s/he would have learned 4 words after Trial One (6 – 2) out of a possible 10 words left to learn (12 words on the list – 2 words learned on Trial One). This would result in an LR score of 0.4 or 40% (i.e., 4 / 10). LR scores range from 0.0 (0% acquired after trial one) to 1.0 (100% acquired), with “normal” performances typically ranging from 50-70% in older adults (Hammers, Spencer, & Apostolova, 2022; Hammers, Suhrie, Dixon, Gradwohl, Duff, et al., 2022). As LR takes the first trial performance into account when considering learning slope, it has been shown to be a superior calculation to other slope metrics (Boscarino, Weitzner, Bailey, Kamper, & Vanderbleek, 2024; Hall et al., 2023).

Similar to the paucity of research on PE in individual learning trials during repeat assessment, PE on learning slopes has not been investigated. This is of particular interest because learning slopes are by definition examining rates of learning across repeated exposure to stimuli *during* a single learning session, therefore they should theoretically be similar to PE – the latter of which is the benefit from repeated exposure *between* testing sessions. Beyond looking at PE for learning slopes, to date learning slopes have very rarely been considered in repeated assessment overall. Hammers and colleagues (Hammers, Spencer, et al., 2022) examined stability of the LR and two other learning slope metrics in robustly cognitively intact older adults over 6-, 12-, and 24-months. Intraclass correlations

ranged from .75 to .79 for LR over that time frame, and from .56 to .60 for the other slope metrics. Although limited, these results suggest that researchers are beginning to explore the clinical utility of learning process scores when assessed over time, with the hope that eventually such metrics may be incorporated into the assessment of cognitive trajectories.

Given the lack of definitive understanding about the impact of PE on individual learning trial performance and learning slopes, the aim of the current research was to investigate these factors in a sample of older adults with either intact cognition or Mild Cognitive Impairment (MCI). It was hypothesized that patients receiving the same versions of the Hopkins Verbal Learning Test – Revised (HVLN-R; Brandt & Benedict, 2001) and the Brief Visuospatial Memory Test – Revised (BVMT-R; Benedict, 1997) twice over a one-week period would experience PE on summary total scores, individual learning trials, and learning slopes. In particular, it was anticipated that the benefit from practice would be most impactful during early learning trials, and progressively diminish in later trials. As such, it was expected that the PE for a measure's total recall score would be primarily driven by improvement at follow-up during the first one to two learning trials. Further, it was hypothesized that these results would hold across diagnostic groups, but that the PE observed for the MCI participants would be reduced relative to those who were cognitively intact. These results are expected to occur despite the potential for interference from ceiling effects – especially in cognitively intact participants. The outcome of these current analyses could shed light on the impact of PE on individual learning trials or learning slopes, which might provide further utility for these process scores in clinical decision making.

METHOD

Participants

Data from 282 community dwelling older adults recruited from the community (e.g., senior centers and independent living facilities) were used in the current analyses. This sample was originally recruited for a study of PE in cognitively normal participants and those with MCI (Duff, Atkinson, et al., 2017). The diagnostic classification procedure for the current sample has been previously described (Duff, Atkinson, et al., 2017; Duff, Dalley, Suhrie, & Hammers, 2019). Briefly, both criteria from Albert and colleagues (Albert et al., 2011) and Petersen (Petersen, 2004) were used to classify participants as cognitively normal (CN) versus having MCI. Application of these criteria incorporated collateral report and a baseline cognitive evaluation (including the Repeatable Battery for the Assessment of Neuropsychological Status [RBANS; Randolph, 2012] and the Trail Making Test (TMT; Reitan, 1992). Cognitive impairment across a domain (e.g., attention, language, memory, executive functioning) was defined as performing 1.0 *SD* below expectation for premorbid intellect (<16th percentile). This translates into a discrepancy of 15.0 *Standard Score* (*SS*) points between Wide Range Achievement Test – 4 Reading (WRAT-4; Wilkinson & Robertson, 2006) performance and the average cognitive domain performance. To ensure cognitive severity, at least one individual task performance within the cognitive domain had to be at least 1.5 *SD* (22.5 *SS* points) below WRAT-4 Reading performance. To aid in understanding, the following example is provided: for a participant with a WRAT-4 Reading score of *SS* = 100, their cognitive domain performance would be classified as impaired if (1)

the average performance for the domain was $< SS = 85$ and (2) a specific task performance within that domain was $< SS = 77.5$. Based on this criteria, 151 participants were classified as CN, and 131 were classified as MCI. Of note, given the focus of the original study, all participants with MCI were of the amnesic subtype (either single-domain or multi-domain).

Inclusion criteria for participants – regardless of diagnosis – included being >64 years of age and functionally independent. Exclusion criteria included a formal diagnosis of dementia, neurological conditions likely to affect cognition (e.g., multiple sclerosis, epilepsy), current severe depression, major psychiatric condition, anti-convulsant or anti-psychotic medications, history of or present substance abuse, and residence in a nursing or skilled living facility.

Procedure

All procedures were approved by local Institutional Review Board before the initiation of the study, with participants providing informed consent prior to completing procedures. The following measures – germane to the current study – were administered at a baseline visit:

- BVMT-R is a measure of visual memory utilizing six geometric designs in six locations on a card, presented across three trials. The Total Recall score is calculated as the number of correctly drawn designs and locations summed across the three trials (range = 0 – 36). Higher values indicate better performance.
- HVLT-R is a measure of verbal memory utilizing 12 words presented over three trials. The Total Recall score is calculated as the number of correctly recalled words summed across the three trials (range = 0 – 36). Higher values indicate better performance.
- RBANS is a neuropsychological test battery containing 12 subtests that are used to calculate Index scores for domains of immediate memory, attention, language, visuospatial/constructional, delayed memory, and global neuropsychological functioning. RBANS index scores utilize age-corrected normative comparisons from the test manual to generate standard scores ($M = 100$, $SD = 15$). Higher scores indicate better performance.
- WRAT-4 Reading subtest is used as an estimate of premorbid intellect where an individual attempts to pronounce irregular words. This score is converted into an age-adjusted SS ($M = 100$, $SD = 15$). Higher values indicate better performance.
- Trail Making Test Parts A and B is a measure of complex mental flexibility and executive functioning, requiring an individual draw a line from either numbers (Part A) or alternating numbers and letters (Part B) in increasing order on a page as quickly as possible (Reitan, 1992). Total seconds to complete the task is the variable of interest for both parts, with lower values indicating better performance.
- The Geriatric Depression Scale (Yesavage et al., 1982) is a 30-item self-report questionnaire used to assess depressive symptoms. Higher scores indicated greater self-reported depression.

After approximately one week ($M = 7.6$ days, $SD = 2.2$), the BVMT-R and HVLTR were repeated to generate a 1 Week PE value. The same form of each test was used to maximize practice effects. The RBANS, WRAT-4, and GDS were only administered at baseline, and participants were classified as CN or MCI based on their performance on baseline scores.

Calculation of Learning Ratio—LR scores were derived from performance on learning trials of both the BVMT-R and the HVLTR. LR reflects the degree of information learned after the first trial on a multi-trial learning task relative to the information left to learn after the first trial. It was calculated as the difference in performance between the final trial and the first trial in the numerator, and the difference between the total points available for a trial and Trial One performance in the denominator (Spencer et al., 2020). The “Total Points Available for a Trial” for both BVMT-R and HVLTR is 12. The specific algorithm is as follows:

$$LR = \frac{(\text{Final Trial performance} - \text{Trial One performance})}{(\text{Total Points Available for a Trial} - \text{Trial One performance})}$$

Calculation of 1 Week PE—In the current study, 1 Week PE were calculated as the difference between Baseline performance and 1 Week performance for the following BVMT-R and HVLTR scores: Total Recall, Trial One, Trial Two, Trial Three, and LR. The equation for PE for each of the relevant learning variables is as follows:

$$PE = 1 \text{ Week performance} - \text{Baseline performance}$$

Of note, the above equation reflects a difference in performance between two timepoints, which may theoretically reflect change due to a multitude of causes. However, this discrepancy was interpreted to be resulting from practice effects based on the presumption that the most likely explanation for improvement over such a short duration was benefit from prior exposure to the test stimuli (and not the initiation of medications or cognitive training). Additionally, while more rigorous calculations for PE have been used in the literature (including standardized regression-based approaches [McSweeney, Naugle, Chelune, & Luders, 1993]), the “Simple Difference Method” (Dikmen, Heaton, Grant, & Temkin, 1999; Duff et al., 2005) was used because it captures raw *actual* changes in scores observed between two time points, without incorporating *expected* change (which is common in regression-based approaches). Larger PE values reflected greater performance at 1 Week relative to Baseline.

Data Analysis—To determine the appropriateness of covariates in the diagnostic group analyses, *independent samples t tests* were conducted between continuous demographic variables (e.g., age, education, retest interval, etc.) and diagnostic group, and *chi-square* analyses were calculated between categorical demographic variables (e.g., sex and ethnicity) and diagnostic group. *Independent samples t tests* compared participants’ baseline cognitive performances on the RBANS between groups, and *one-sample t tests* examined the presence of 1 Week PE.

Differences in Learning Trial PE and LR PE Across and Between Groups—

Comparison of HVLTR and BVMT-R learning trial PE performance both between and within diagnostic groups was undertaken using a series of *two-way (between-within) repeated measures MANCOVA*. The main effects of trial PE (*within*; e.g., Trial One PE versus Trial Two PE in CN samples) and diagnostic group (*between*) were examined on the dependent variables of BVMT-R Total Recall PE and HVLTR Total Recall PE, and a trial PE x diagnostic group interaction effect was also included. Following significant omnibus testing, *ANCOVA* and *independent samples t tests* were conducted to determine specific differences between trial PE.

Additionally, to determine whether 1 Week PE was diminished in MCI groups relative to CN for a specific learning trial, Total Recall, or LR, *one-way MANCOVA* was performed, with subsequent *ANCOVA* used to identify the specific variable-differences following significant omnibus testing. While similar sounding, this *one-way MANCOVA* answers a different question than the *two-way repeated measures MANCOVA* described above. Whereas the *two-way repeated measures MANCOVA* has the ability to examine differences in PE between diagnostic groups, it can only do so when collapsing performance across trials (e.g., PE for Trial One + Trial Two + Trial Three for CN versus PE for Trial One + Trial Two + Trial Three for MCI). Consequently, the only way to examine diagnostic group differences for a specific learning trial (e.g., Trial Two PE between CN and MCI), or for LR, is using these secondary *one-way MANCOVA* analyses.

Supplementary Analysis—To understand the unique contribution of learning trial PE on the prediction of the PE for Total Recall for the interested reader, a series of supplementary *hierarchical linear regression* analyses can be found in the **Appendix**. Specifically, predictive demographic variables were included in Step 1, Trial One PE was in Step 2, Trial Two PE was in Step 3, and Trial Three PE was in Step 4. These *hierarchical linear regression* models predicted PE on Total Recall from the BVMT-R and HVLTR, in CN and MCI samples separately.

Measures of effect size were expressed as Cohen's *d* (*t tests and MANCOVA*), r^2 values (*hierarchical linear regression*), and ϕ (Phi; *chi-square*). For effect size comparisons for individual trial PE across diagnostic groups (e.g., the magnitude of Trial One PE's effect for the CN group relative to the MCI group), 95% *Compatibility Intervals (CIs)* were calculated for the *Cohen's d* values; significance was determined by failure of one variable's 95% *CI* to overlap with the mid-point of the other variable's 95% *CI* (Cumming & Finch, 2005). To protect against multiple comparisons, a Holm-Bonferroni method of adjustment of the two-tailed alpha level was undertaken for all primary analyses.

RESULTS

Preliminary Analyses

Table 1 reflects demographic characteristics of participants from the current study's CN and MCI samples. The CN group was significantly younger and had a lower premorbid intellect than the MCI group, $ps = .001$ to $.003$, $ds = -0.36$ to -0.53 . No differences were observed between groups for education, sex, ethnicity, retest interval, or self-reported depression,

$ps = .18$ to $.85$, $ds = -0.11$ to 0.11 , $Phis = .05$ to $.08$. Consistent with their diagnostic make-up, the CN group performed statistically better than the MCI group on tasks from the RBANS pertaining to Immediate Memory, $p < .001$, $d = 0.78$, Language, $p = .002$, $d = 0.37$, Attention, $p = .001$, $d = 0.31$, Delayed Memory, $p < .001$, $d = 1.16$, and Total Scale score, $p < .001$, $d = 0.86$. No difference was observed for RBANS Visuospatial/Constructional, $p = .19$, $d = 0.16$. Consequently, age was used as a covariate for analyses comparing PE between groups. Premorbid intellect was not used as a covariate given its role in the diagnostic classification process.

Table 2 shows the mean Baseline and 1 Week performances for learning trial and LR for the BVMT-R and HVLTR for CN participants, along with 1 Week PE values. As a reminder, higher PE scores reflect stronger performance (or larger learning slope) at 1 Week follow-up. Table 3 shows the performances for MCI participants. In general, at Baseline the majority of information learned on the HVLTR and BVMT-R occurred during Trial One, with less information gained at subsequent trials. For example, CN participants learned on average 7.3 items on Trial One for the HVLTR, 9.6 items (a gain of 2.3) on Trial Two, and 10.5 items (a gain of 0.9) on Trial Three. Additionally, significant PE were present for both diagnostic groups across Total Recall and learning trials for BVMT-R and HVLTR, $ps < .001$, $ds = 0.35$ to 1.98 . No PE were observed for LR, $ps = .28$ to $.84$, $ds = -0.02$ to 0.10 .

Differences in Learning Trial PE and LR PE Across and Between Groups

Two-way repeated measures MANCOVAs were undertaken to examine differences in learning trial 1 Week PE for both BVMT-R and HVLTR between diagnostic groups and within trials, after controlling for age. Significant differences were observed in the omnibus test for both BVMT-R and HVLTR, $ps = .001$ to $.03$, $ds = 0.31$ to 0.67 . Specifically, both BVMT-R and HVLTR possessed a significant learning trial PE x diagnostic group interaction effect, $ps = .001$ to $.015$, $ds = 0.26$ to 0.54 . Additionally, main effects for trial PE and diagnostic group were observed for BVMT-R ($p = .002$, $d = 0.31$ and $p = .008$, $d = 0.32$, respectively), though no main effects were observed for HVLTR ($p = .10$, $d = 0.26$ and $p = .38$, $d = 0.11$, respectively). As the interaction effect was significant, we will focus on that finding. Specifically, the interaction effects can be explained by post-hoc *independent samples t tests* revealing different relationships between learning trials PE depending on the diagnostic group. As can be observed in Figure 1, for the CN group the PE for BVMT-R Trial One was larger than that for Trial Two, which was larger than that for Trial Three ($ps < .001$, $ds = 0.54$ to 0.81). Conversely for the MCI group, the PE for BVMT-R Trial One was larger than that for Trial Two ($p = .003$, $d = 0.27$), but no difference in PE existed between Trial Two and Trial Three ($p = .24$, $d = 0.10$). A similar effect can be observed in Figure 1 for HVLTR. Overall across both BVMT-R and HVLTR, there was a significantly larger discrepancy for PE between Trial One and Trial Two for the CN group than the MCI group (BVMT-R: $d = 0.81$ [95% $CI = 0.57 - 1.05$] for CN vs $d = 0.27$ [95% $CI = 0.03 - 0.51$] for MCI; HVLTR: $d = 0.59$ [95% $CI = 0.35 - 0.83$] for CN vs $d = 0.23$ [95% $CI = 0.00 - 0.46$] for MCI). Stated another way, the difference in Trial PE between groups appears to be driven by larger PE for Trial One relative to the other Trials for the CN group compared to the MCI group.

Finally, Figure 2 displays whether 1 Week PE was diminished in MCI groups relative to CN for a specific learning trial, Total Recall, or LR (e.g., Trial 2 PE for CN versus Trial 2 PE for MCI) using *one-way MANCOVA* analyses. Note, Figure 2 represents comparable raw data values to Figure 1, but presented to reflect CN versus MCI comparisons. Following significant omnibus tests for both BVMT-R and HVLTR (ps = .001 to .005, ds = 0.56 to 0.78), *ANCOVA* results revealed that PE were significantly larger for the CN group relative to the MCI group for BVMT-R Total Recall, $p = .009$, $d = 0.31$, and BVMT-R Trial One, $p < .001$, $d = 0.65$. No differences in PE were observed between groups for BVMT-R Trial Two, Trial Three, LR, and HVLTR Total Recall, Trial One, Trial Two, Trial Three, or LR (all ps > .05).

DISCUSSION

The current study observed that 1 Week PE were present for the BVMT-R and the HVLTR for both summary total scores and individual learning trials. These findings are consistent with a multitude of past research suggesting that benefit from practice is commonly observed in total scores for many memory measures (Calamia et al., 2012; Duff & Hammers, 2022), including the BVMT-R and HVLTR (Duff et al., 2018; Duff, Hammers, et al., 2017; Duff, Suhrie, et al., 2019). As can be observed in Table 2, for example, the effect sizes for the Total Recall score PE in CN participants were large ($ds = 1.12 - 1.98$), suggesting that participants improved – on average – 9 points on the BVMT-R (out of a possible 36 points) during the second administration a week later, and 4 points on the HVLTR. These results represent the first documentation of observed PE in individual learning trials, and suggest that PE across these measures are not limited by ceiling effects in these CN and MCI samples. As can be seen in Tables 2 and 3, PE were present across all three learning trials for both memory measures – for both CN and MCI participants. Examining the results of BVMT-R for CN participants in Table 2 more specifically, we can observe that Trial One performance improved from acquiring approximately 3.5 out of 12 words on average at Baseline to a little over 8 out of 12 words at 1 Week. These magnitudes of effect appear to rival (e.g., $ds = 1.93$ vs. 1.98 ; BVMT-R) or exceed (e.g., $ds = 1.29$ vs. 1.12 ; HVLTR) those for the Total Recall score PE.

The severity of this improvement at both the Total Recall and Trial One levels has practical implications for clinicians. For example, when applying normative comparisons provided by the test developer for the BVMT-R (Benedict, 1997), this doubling of Trial One performance as a result of previous exposure would result in an improvement from a *T* score of 44.5 (30thile) to a *T* score of 68.5 (97thile) for a 75-year-old individual. Similarly, using the PE scores for the MCI sample (Table 3), our hypothetical 75-year-old patient would be re-classified as being cognitively normal based on these 1 Week performances (improvement from the borderline impaired range to a cognitively average performance). These examples speak to the previously-suggested potential for PE to influence diagnostic decision making (Sanderson-Cimino et al., 2022), as distinct clinical pictures can currently be seen depending on whether the presence of PE was considered. Because serial assessment will likely be increasingly used for diagnostic monitoring of disease modifying treatments for AD – especially after evidence of amyloid-related imaging abnormality (ARIA; Cummings et al.,

2023) – clinical neuropsychologists and cognitive neurologists will need to be attuned even more to how benefit from practice could impact the results obtained in clinic.

When examining Trials Two and Three for each measure in these same Tables, we can observe that while significant 1 Week PE are present, the magnitudes of effect are diminished (d s of 0.66 and 0.35 for HVLTR in CN) compared to Trial One PE (d of 1.29). These findings are bolstered by our *hierarchical regression* results in the Appendix suggesting that the benefit from practice observed at Trial One is consistently the strongest contributor of the overall Total Recall PE for a measure (see **Tables S1** and **S2**). Put another way, the collection of findings suggests that majority of the benefit that a participant receives from serial memory assessment appears to present in the first trial of the follow-up assessment. These results support our hypothesis, and can be explained as follows. For a 3-trial learning task like either the HVLTR and BVMT-R, participants receive the same stimuli three times per testing session. As a result, at follow-up, participants have had three opportunities to develop scaffolding (Fiechter & Benjamin, 2019) to learn the information prior to their Trial One presentation at 1 Week; this leads to a substantial increase in acquisition of the information relative to Trial One at Baseline, when the participants had previously never been exposed to the material. By Trials Two and Three at Baseline, participants are already receiving some benefit from *within*-session previous exposure, therefore the resultant improvement on Trials Two and Three at 1 Week (which represent exposures number five and six of the material over a period of one week) are smaller in comparison.

These results coincide with patterns of learning traditionally seen at baseline for multi-trial learning tasks. There exists a tendency toward non-linear learning across trials on many learning tasks, which is exemplified by our HVLTR results at Baseline (Table 2) indicating that the majority of information learned occurred during Trial One, with diminished learning on subsequent trials. This suggests that clinicians and researchers may want to focus on Trial One performance as being a purer assessment of learning acquisition – associated with cortical connectivity of the dorsal attention network (Putchá, Brickhouse, Wolk, & Dickerson, 2019) – and that future research should consider Trial One PE as a cognitive biomarker for neurodegenerative disease.

Our interpretation of PE related to learning slopes was more challenging than that for Total Recall and Trial One. This line of investigation was relevant given that learning slopes are themselves related to learning across repeated exposure to stimuli *within* a single learning session (Walhovd et al., 2020). LR failed to display an appreciable PE, and was characterized by both non-significant findings (p s > .05) and negligible effect sizes (d s = -0.02 to 0.10). As this is the first investigation of practice effects in learning slopes, there is no prior literature in which to compare these findings. These slope results are counter to hypotheses and expectations, and likely arise from the method in which the learning slopes are calculated. As has been consistently shown by our individual trial results above, overall improvement at 1 Week for the Total Recall scores tended to be driven by enhanced Trial One performance. This improved Trial One performance at follow-up means that there is inherently less information available to learn over Trials Two and Three. As will be described in more detail below, this appears to lead to 1) a restriction of range of LR

values as a result of ceiling effects and 2) an over-penalization of missing items for LR and potential for negative values.

First, because the maximum value for LR cannot exceed 1.0 (or 100%), a ceiling effect is created whereby LR scores have limited ability to improve at follow-up regardless of the presence of PE. This subsequently truncates the scoring distribution for LR, leading to a restriction of range. Second, as “information left to learn after Trial One” is the denominator of the LR equation, improved Trial One performance at follow-up will lead to a smaller denominator for the LR equation. This smaller denominator will result in a greater penalty for each item not learned. An example will help explain this idea further. Let’s suppose that on HVLTR at Baseline a participant receives a Trial One score of 4 of 12 and a score of 10 of 12 for Trial Three; this results in an LR score of 0.75 (or 75%; $10 - 4 / 8$ words left to learn). If this same participant subsequently performs better on Trial One at follow-up with a score of 8 of 12, but still scores 10 of 12 at Trial Three, the resultant LR score is 0.50 (or 50%; $10 - 8 / 4$ words left to learn). When incorporated into our PE equation, we get $0.50 - 0.75$, so this participant appears to have performed worse at follow-up even though their Trial One performance improved between testing sessions. As a result, although the clinical utility of learning slopes has repeatedly been supported (see Spencer et al., 2023 for a recent systematic review), PE in LR appears to be mathematically confounded. Consideration of PE in learning slopes consequently does not appear to be advised in clinical practice. Viewed independently at each time point, however, these results highlight that LR seems to be a relatively stable metric across testing sessions. As such, LR may not be susceptible to pronounced changes as a result of repeated exposure to stimuli, which further supports its use at baseline clinically and in research settings.

When examining differences in PE between diagnostic groups, a few notable findings arose. First, results from our *two-way repeated measures MANCOVA* interactions suggest that across both measures, the importance of Trial One PE towards overall learning was greater in CN participants relative to those with MCI. For example, Figure 1a shows a large step-down in effect for CN participants between Trials One, Two, and Three; conversely, MCI participants appeared to improve across trials more equally as a result of practice. This indicates that while participants with MCI are still capable of benefiting from PE, they do not experience the large gain in performance at Trial One of follow-up seen in CN participants. Instead, they appear to require successive trials (i.e., Trials Two and Three) to gradually improve their performance. This diminished-PE finding in MCI is similar to the observation in other research that the primacy effect (remembering more items from the beginning of a list-learning task) is commonly reduced in amnesic populations (Foldi, Brickman, Schaefer, & Knutelska, 2003) – and especially in those with hippocampal atrophy (Gicas et al., 2020). As worse PE on Total Recall scores have been associated with decreased hippocampal volumes (Duff et al., 2018; Duff, Suhrie, et al., 2019), there are hints that these cognitive processes may be driven by similar psychological mechanisms and neural structures, which will require future examination to consider in greater depth. Relatedly, as these findings raise the possibility of diagnostic utility of Trial One performance, future research is needed to investigate whether the amount of information learned during Trial One of repeated assessments is as sensitive or more sensitive an indicator of MCI or AD pathology than what has been observed thus far with Total Recall scores.

A second notable finding is that while the CN group displayed stronger PE for BVMT-R Total score and Trial One than the MCI group, no differences between CN and MCI were seen across trials or Total score for HVLTR (Figure 2). These results are somewhat surprising given the literature on the limited capacity of those with MCI and/or AD to benefit from practice. As summarized in a recent systematic review, Jutten and colleagues (2020) found consistent support for the hypothesis that PE tend to be smaller in participants with MCI than those with CN. This effect is often associated with the notion of “the rich get richer”, such that higher baseline performance bestows greater opportunity to benefit from prior exposure of the material (Patton et al., 2005). However, this finding is not universal, such that Duff et al. (2008) has observed that participants with MCI displayed stronger PE on the HVLTR than cognitively intact groups. As all MCI presentations are not created equally – as a function of either subtype presentation or underlying etiology – it is possible that our currently discrepant findings may be a consequence of our MCI sample. For example, Hammers and colleagues (2022) observed that amnesic MCI participants recruited from a memory disorders clinic displayed worse baseline performances and smaller long-term PE relative to an amnesic MCI sample recruited from the community. As the current MCI sample was all community-recruited, it is possible that the disease severity in our MCI participants was not large enough to result in diminished PE across both BVMT-R and HVLTR measures. Given our ambiguous findings, future investigation is warranted to better understand the influences of PE in MCI.

Some limitations existed to the current study. First, these results are specific to the HVLTR and BVMT-R over a retest interval of one week, as well as to the LR metric derived using the equations from Spencer and colleagues (2020). As a result, generalizability cannot be automatically assumed across different cognitive measures and learning slope equations, retest intervals, or test versions. This point is particularly true for the retest interval chosen; while the one-week retest interval was selected to maximize PE, this is not a common interval used clinically. Second, we cannot speak to the usability (or lack thereof) of PE across all learning slopes, but only when using LR. Third, these results may not generalize to more heterogeneous participants as it pertains to premorbid functioning, sex, education, and ethnicity, especially given that only one participant in our sample was non-White. Future research should consider PE in individual learning trials in samples that are not mostly well-educated, female, and White. Fourth, as alluded to in the Methods, the current study used the Simple Difference Method to calculate one-week PE (i.e., 1 Week score – Baseline score), despite the availability of more sophisticated methods (e.g., standardized regression-based approaches). This was done to capture the “purest” measure of raw change between two testing sessions during this first documentation of PE in individual learning trials, without accounting for the degree of change that was predicted based on demographic and clinical variables, and Time 1 score. Future work that builds on these introductory findings may benefit from incorporating more advanced PE indexes. Finally, consistent with previous research (Hammers, Suhrie, Dixon, Gradwohl, Archibald, et al., 2022; Spencer et al., 2020), an LR value of 1.00 (100% learned) was assigned in the scenario when participants learned all available items on Trial One on either measure at each time point. This was necessary because a score of 12 of 12 on Trial One would result in a value of zero in the denominator of LR (Total Points Available For a Trial – Trial One), which

would result in an undefined mathematical expression. While a “perfect” score on Trial One is relatively uncommon in most learning slope studies, the frequency of this occurrence increased somewhat in the current study (29 cases for HVLt-R, 10 cases for BVMT-R) given that the greatest improvement in performance as a result of PE occurred at Trial One of the 1 Week follow-up. Despite these limitations, PE on learning trials and learning slopes provide some similarities and differences with the broader literature on PE, but remain an area in need of additional investigation. As a consequence, clinicians and researchers looking to incorporate PE into their diagnostic considerations would be encouraged to do so for learning trials but not learning slopes, as the former appears to be less susceptible to ceiling effects.

Funding and Acknowledgements:

Data collection for this project was supported by an anonymous foundation and research grants from the National Institutes on Aging (K23 AG028417 and 5R01 AG055428).

REFERENCES

- Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, ... Phelps CH (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 7(3), 270–279. doi:10.1016/j.jalz.2011.03.008
- Benedict R. (1997). *Brief visuospatial memory test - Revised: professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.
- Boscarino JJ, Weitzner DS, Bailey EK, Kamper JE, & Vanderbleek EN (2024). Utility of learning ratio scores from the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Word List Memory Test in distinguishing patterns of cognitive decline in veterans referred for neuropsychological evaluation. *Clin Neuropsychol*, 1–13. doi:10.1080/13854046.2024.2330144
- Brandt J, & Benedict R (2001). *Hopkins Verbal Learning Test - Revised*. Odessa, FL: PAR.
- Bubu OM, Andrade AG, Umasabor-Bubu OQ, Hogan MM, Turner AD, de Leon MJ, ... Osorio RS (2020). Obstructive sleep apnea, cognition and Alzheimer's disease: A systematic review integrating three decades of multidisciplinary research. *Sleep Med Rev*, 50, 101250. doi:10.1016/j.smrv.2019.101250 [PubMed: 31881487]
- Calamia M, Markon K, & Tranel D (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *Clin Neuropsychol*, 26(4), 543–570. doi:10.1080/13854046.2012.680913 [PubMed: 22540222]
- Chelune G, & Duff K (2012). *The assessment of change: Serial assessments in dementia evaluation*. New York, NY: Springer.
- Cohen CI, Reisberg B, & Yaffee R (2022). Global cognitive trajectory patterns in Alzheimer's disease. *Int Psychogeriatr*, 1–10. doi:10.1017/s1041610222000047
- Cumming G, & Finch S (2005). Inference by eye: confidence intervals and how to read pictures of data. *Am Psychol*, 60(2), 170–180. doi:10.1037/0003-066x.60.2.170 [PubMed: 15740449]
- Cummings J, Apostolova L, Rabinovici GD, Atri A, Aisen P, Greenberg S, ... Salloway S (2023). Lecanemab: Appropriate Use Recommendations. *J Prev Alzheimers Dis*, 10(3), 362–377. doi:10.14283/jpad.2023.30 [PubMed: 37357276]
- das Nair R, Cogger H, Worthington E, & Lincoln NB (2016). Cognitive rehabilitation for memory deficits after stroke. *Cochrane Database Syst Rev*, 9, CD002293. doi:10.1002/14651858.CD002293.pub3 [PubMed: 27581994]
- Dikmen SS, Heaton RK, Grant I, & Temkin NR (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *J Int Neuropsychol Soc*, 5(4), 346–356. [PubMed: 10349297]

- Duff K, Anderson JS, Mallik AK, Suhrie KR, Atkinson TJ, Dalley BCA, ... Hoffman JM (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 [PubMed: 30143414]
- Duff K, Atkinson TJ, Suhrie KR, Dalley BC, Schaefer SY, & Hammers DB (2017). Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. *J Clin Exp Neuropsychol*, 39(4), 396–407. doi:10.1080/13803395.2016.1230596 [PubMed: 27646966]
- Duff K, Beglinger LJ, Moser DJ, Schultz SK, & Paulsen JS (2010). Practice effects and outcome of cognitive training: preliminary evidence from a memory training course. *Am J Geriatr Psychiatry*, 18(1), 91. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20104658> [PubMed: 20104658]
- Duff K, Beglinger LJ, Schoenberg MR, Patton DE, Mold J, Scott JG, & Adams RL (2005). Test-retest stability and practice effects of the RBANS in a community dwelling elderly sample. *J Clin Exp Neuropsychol*, 27(5), 565–575. doi:10.1080/13803390490918363 [PubMed: 16019633]
- Duff K, Beglinger LJ, Van Der Heiden S, Moser DJ, Arndt S, Schultz SK, & Paulsen JS (2008). Short-term practice effects in amnesic mild cognitive impairment: implications for diagnosis and treatment. *Int Psychogeriatr*, 20(5), 986–999. doi:10.1017/S1041610208007254 [PubMed: 18405398]
- Duff K, Dalley BCA, Suhrie KR, & Hammers DB (2019). Predicting Premorbid Scores on the Repeatable Battery for the Assessment of Neuropsychological Status and their Validation in an Elderly Sample. *Arch Clin Neuropsychol*, 34(3), 395–402. doi:10.1093/arclin/acy050 [PubMed: 29878036]
- Duff K, & Hammers D (2022). Practice effects in mild cognitive impairment: A validation of calamia et al. (2012). *Clin Neuropsychol*, 36(3), 571–583. doi:10.1080/13854046.2020.1781933 [PubMed: 32594886]
- Duff K, Hammers DB, Dalley BCA, Suhrie KR, Atkinson TJ, Rasmussen KM, ... Hoffman JM (2017). Short-Term Practice Effects and Amyloid Deposition: Providing Information Above and Beyond Baseline Cognition. *J Prev Alzheimers Dis*, 4(2), 87–92. doi:10.14283/jpad.2017.9 [PubMed: 28966919]
- Duff K, Horn KP, Foster NL, & Hoffman JM (2015). Short-Term Practice Effects and Brain Hypometabolism: Preliminary Data from an FDG PET Study. *Arch Clin Neuropsychol*, 30(3), 264–270. doi:10.1093/arclin/acv018 [PubMed: 25908614]
- Duff K, Suhrie KR, Dalley BCA, Anderson JS, & Hoffman JM (2019). External validation of change formulae in neuropsychology with neuroimaging biomarkers: A methodological recommendation and preliminary clinical data. *Clin Neuropsychol*, 33(3), 478–489. doi:10.1080/13854046.2018.1484518 [PubMed: 29884099]
- Fiechter JL, & Benjamin AS (2019). Techniques for scaffolding retrieval practice: The costs and benefits of adaptive versus diminishing cues. *Psychon Bull Rev*, 26(5), 1666–1674. doi:10.3758/s13423-019-01617-6 [PubMed: 31161529]
- Foldi NS, Brickman AM, Schaefer LA, & Knutelska ME (2003). Distinct serial position profiles and neuropsychological measures differentiate late life depression from normal aging and Alzheimer's disease. *Psychiatry Res*, 120(1), 71–84. doi:10.1016/s0165-1781(03)00163-x [PubMed: 14500116]
- Gagnon C, Saillant K, Olmand M, Gayda M, Nigam A, Bouabdallaoui N, ... Bherer L (2022). Performances on the Montreal Cognitive Assessment Along the Cardiovascular Disease Continuum. *Arch Clin Neuropsychol*, 37(1), 117–124. doi:10.1093/arclin/acab029 [PubMed: 33960374]
- Gicas KM, Honer WG, Wilson RS, Boyle PA, Leurgans SE, Schneider JA, & Bennett DA (2020). Association of serial position scores on memory tests and hippocampal-related neuropathologic outcomes. *Neurology*, 95(24), e3303–e3312. doi:10.1212/wnl.0000000000010952 [PubMed: 33144516]
- Gifford KA, Phillips JS, Samuels LR, Lane EM, Bell SP, Liu D, ... Jefferson AL (2015). Associations between Verbal Learning Slope and Neuroimaging Markers across the Cognitive Aging Spectrum. *J Int Neuropsychol Soc*, 21(6), 455–467. doi:10.1017/s1355617715000430 [PubMed: 26219209]
- Hall MG, Wollman SC, Haines ME, Boyle MA, Richardson HK, & Hammers DB (2023). Novel learning ratio from the NAB list learning test distinguishes between clinical groups: clinical validation and sex-related differences. *J Clin Exp Neuropsychol*, 1–12. doi:10.1080/13803395.2023.2236772 [PubMed: 37083506]

- Hammers DB, Kostadinova RV, Spencer RJ, Ikanga JN, Unverzagt FW, Risacher SL, & Apostolova LG (2023). Sensitivity of memory subtests and learning slopes from the ADAS-Cog to distinguish along the continuum of the NIA-AA Research Framework for Alzheimer's Disease. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 30(6), 866–884. doi:10.1080/13825585.2022.2120957 [PubMed: 36074015]
- Hammers DB, Lin JH, Polsinelli AJ, Logan PE, Risacher SL, Schwarz AJ, & Apostolova LG (2023). Criterion Validation of Tau PET Staging Schemes in Relation to Cognitive Outcomes. *J Alzheimers Dis*, 96(1), 197–214. doi:10.3233/jad-230512 [PubMed: 37742649]
- Hammers DB, Nemes S, Diedrich T, Eloyan A, Kirby K, Aisen P, ... Apostolova LG (2023). Learning slopes in early-onset Alzheimer's disease. *Alzheimers Dement*, 19 Suppl 9, S19–s28. doi:10.1002/alz.13159 [PubMed: 37243937]
- Hammers DB, Pentchev JV, Kim HJ, Spencer RJ, & Apostolova LG (2023). The relationship between learning slopes and Alzheimer's Disease biomarkers in cognitively unimpaired participants with and without subjective memory concerns. *J Clin Exp Neuropsychol*, 1–17. doi:10.1080/13803395.2023.2254444 [PubMed: 37083506]
- Hammers DB, Spencer RJ, & Apostolova LG (2022). Validation of and Demographically Adjusted Normative Data for the Learning Ratio Derived from the RAVLT in Robustly Intact Older Adults. *Arch Clin Neuropsychol*, 37(5), 981–993. doi:10.1093/arclin/acac002 [PubMed: 35175287]
- Hammers DB, Suhrie K, Dixon A, Gradwohl BD, Archibald ZG, King JB, ... Hoffman JM (2022). Relationship between a novel learning slope metric and Alzheimer's disease biomarkers. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn*, 29(5), 799–819. doi:10.1080/13825585.2021.1919984 [PubMed: 33952156]
- Hammers DB, Suhrie K, Dixon A, Gradwohl BD, Duff K, & Spencer RJ (2022). Validation of HVLt-R, BVMT-R, and RBANS Learning Slope Scores along the Alzheimer's Continuum. *Arch Clin Neuropsychol*, 37(1), 78–90. doi:10.1093/arclin/acab023 [PubMed: 33899087]
- Hammers DB, Suhrie KR, Porter SM, Dixon AM, & Duff K (2022). Validation of one-year reliable change in the RBANS for community-dwelling older adults with amnesic mild cognitive impairment. *Clin Neuropsychol*, 36(6), 1304–1327. doi:10.1080/13854046.2020.1807058 [PubMed: 32819188]
- Hinton-Bayre AD (2016). Clarifying Discrepancies in Responsiveness Between Reliable Change Indices. *Arch Clin Neuropsychol*, 31(7), 754–768. doi:10.1093/arclin/acw064 [PubMed: 27590303]
- Holm SP, Wolfer AM, Pointeau GHS, Lipsmeier F, & Lindemann M (2022). Practice effects in performance outcome measures in patients living with neurologic disorders - A systematic review. *Heliyon*, 8(8), e10259. doi:10.1016/j.heliyon.2022.e10259 [PubMed: 36082322]
- Jutten RJ, Grandoit E, Foldi NS, Sikkes SAM, Jones RN, Choi SE, ... Rabin LA (2020). Lower practice effects as a marker of cognitive performance and dementia risk: A literature review. *Alzheimers Dement (Amst)*, 12(1), e12055. doi:10.1002/dad2.12055 [PubMed: 32671181]
- McSweeney A, Naugle RI, Chelune GJ, & Luders H (1993). "T-scores for change:" An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist* 7, 300–312.
- Patton DE, Duff K, Schoenberg MR, Mold J, Scott JG, & Adams RL (2005). Base rates of longitudinal RBANS discrepancies at one- and two-year intervals in community-dwelling older adults. *Clin Neuropsychol*, 19(1), 27–44. doi:10.1080/13854040490888477 [PubMed: 15814476]
- Petersen RC (2004). Mild cognitive impairment as a diagnostic entity. *J Intern Med*, 256(3), 183–194. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15324362> [PubMed: 15324362]
- Putchá D, Brickhouse M, Wolk DA, & Dickerson BC (2019). Fractionating the Rey Auditory Verbal Learning Test: Distinct roles of large-scale cortical networks in prodromal Alzheimer's disease. *Neuropsychologia*, 129, 83–92. doi:10.1016/j.neuropsychologia.2019.03.015 [PubMed: 30930301]
- Randolph C. (2012). *Repeatable Battery for the Assessment of Neuropsychological Status*. Bloomington, MN: The Psychological Corporation.
- Reitan R. (1992). *Trail Making Test: Manual for administration and scoring*. Tucson, AZ: Reitan Neuropsychology Laboratory.

- Sanderson-Cimino M, Elman JA, Tu XM, Gross AL, Panizzon MS, Gustavson DE, ... Alzheimer's Disease Neuroimaging, I. (2022). Cognitive practice effects delay diagnosis of MCI: Implications for clinical trials. *Alzheimer's & dementia (New York, N. Y.)*, 8(1), e12228. doi:10.1002/trc2.12228. (Accession No. 35128027)
- Spencer RJ, Gradwohl BD, Williams TF, Kordovski VM, & Hammers DB (2020). Developing learning slope scores for the repeatable battery for the assessment of neuropsychological status. *Appl Neuropsychol Adult*, 1–7. doi:10.1080/23279095.2020.1791870 [PubMed: 29617165]
- Spencer RJ, Williams TF, Kordovski VM, Patrick SD, Lengu K, Gradwohl BD, & Hammers DB (2023). A quantitative review of competing learning slope metrics: effects of age, sex, and clinical diagnosis. *J Clin Exp Neuropsychol*, 45(7), 744–757. doi:10.1080/13803395.2024.2314741 [PubMed: 38357915]
- Walhovd KB, Bråthen ACS, Panizzon MS, Mowinckel AM, Sørensen Ø, de Lange A-MG, ... Fjell AM (2020). Within-session verbal learning slope is predictive of lifespan delayed recall, hippocampal volume, and memory training benefit, and is heritable. *Scientific Reports*, 10(1), 21158. doi:10.1038/s41598-020-78225-1 [PubMed: 33273630]
- Walhovd KB, Bråthen ACS, Panizzon MS, Mowinckel AM, Sørensen Ø, de Lange AG, ... Fjell AM (2020). Within-session verbal learning slope is predictive of lifespan delayed recall, hippocampal volume, and memory training benefit, and is heritable. *Sci Rep*, 10(1), 21158. doi:10.1038/s41598-020-78225-1 [PubMed: 33273630]
- Warschausky S, Kay JB, Chi P, & Donders J (2005). Hierarchical linear modeling of California Verbal Learning Test--Children's Version learning curve characteristics following childhood traumatic head injury. *Neuropsychology*, 19(2), 193–198. doi:10.1037/0894-4105.19.2.193 [PubMed: 15769203]
- Wilkinson GS, & Robertson GJ (2006). *WRAT 4: Wide Range Achievement Test, professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.
- Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, & Leirer VO (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*, 17(1), 37–49. doi:10.1016/0022-3956(82)90033-4 [PubMed: 7183759]

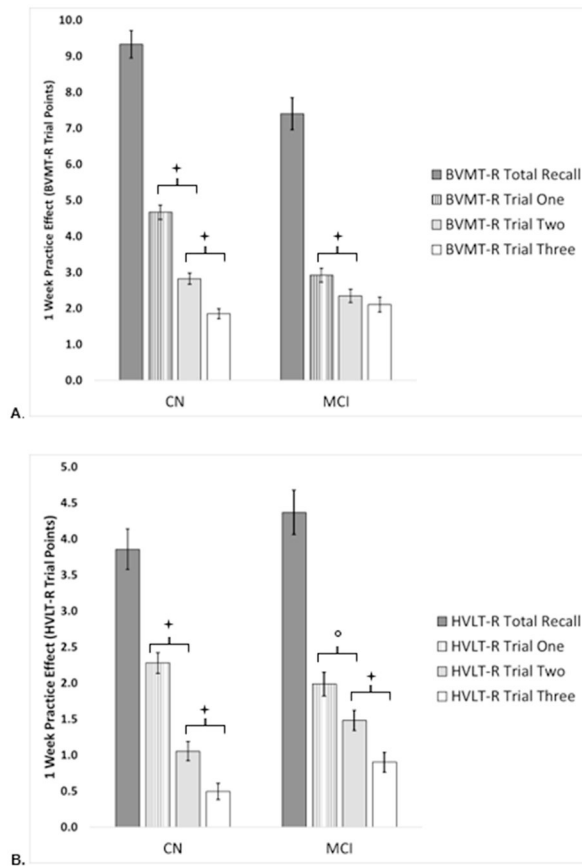


Figure 1. One week Practice Effects for Total Recall, Trial One, Trial Two, and Trial Three of the BVMT-R (A) and the HVLTR-R (B). BVMT-R = Brief Visual Memory Test – Revised, HVLTR-R = Hopkins Verbal Learning Test – Revised, CN = Cognitive Normal, MCI = Mild Cognitive Impairment. Note that the Practice Effect is calculated as 1 Week score – Baseline score. + Denotes significant difference between Trial Practice Effects, $p < .001$. ° Denotes significant difference between Trial Practice Effects, $p < .01$.

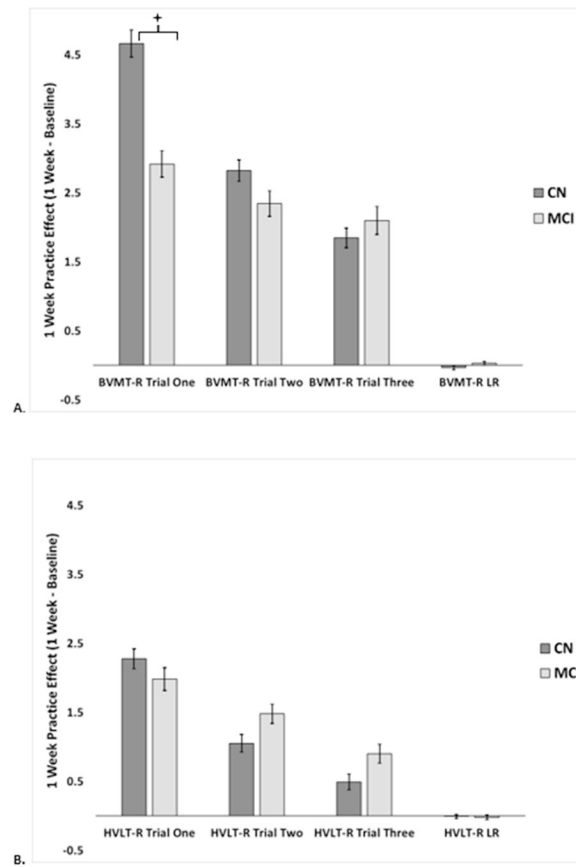


Figure 2.

One week Practice Effects for Trial One, Trial Two, Trial Three, and LR of the BVMT-R (A) and the HVLTR-R (B). BVMT-R = Brief Visual Memory Test – Revised, HVLTR-R = Hopkins Verbal Learning Test – Revised, LR = Learning Ratio, CN = Cognitive Normal, MCI = Mild Cognitive Impairment. Note that the Practice Effect is calculated as 1 Week score – Baseline score. + Denotes significant difference between CN and MCI groups, $p < .001$.

Table 1.

Demographic, neuropsychological, and behavioral variables for the diagnostic groups and total sample.

Variable	Cognitively Normal	Mild Cognitive Impairment	Total Sample
<i>n</i>	151	131	282
Age (years) ¹	75.28 (7.2)	79.13 (7.5)	77.07 (7.6)
Education (years)	15.40 (2.6)	15.34 (2.5)	15.38 (2.6)
Sex (% female)	82.8%	76.3%	79.8%
Minority Status (% Non-Hispanic/ White)	99.3%	100.0%	99.6%
Geriatric Depression Scale	4.06 (3.6)	4.47 (3.5)	4.25 (3.6)
Retest Interval	7.44 (1.6)	7.69 (2.8)	7.56 (2.2)
WRAT-4 Reading ¹	107.54 (7.6)	110.48 (8.7)	108.91 (8.2)
RBANS Immediate Memory Index ¹	109.03 (14.6)	97.35 (15.6)	103.60 (16.1)
RBANS Visuospatial/Constructional Index	103.75 (15.5)	101.26 (16.4)	102.60 (16.0)
RBANS Language Index ²	104.19 (10.8)	99.87 (12.4)	102.18 (11.8)
RBANS Attention Index ³	104.68 (14.7)	100.23 (14.1)	102.61 (14.6)
RBANS Delayed Memory Index ¹	108.32 (9.2)	94.56 (14.4)	101.93 (13.7)
RBANS Total Scale ¹	108.54 (12.5)	97.80 (12.4)	103.55 (13.6)

Note: WRAT-4 = Wide Range Achievement Test – 4, RBANS = Repeatable Battery for the Assessment of Neuropsychological Status, Retest Interval = Time between baseline and one-week evaluations. All values are *Mean (Standard Deviation; range)* unless listed otherwise. ¹ Denotes significant difference between groups, $p < .001$. ² Denotes significant difference between groups, $p < .01$. ³ Denotes significant difference between groups, $p < .05$.

Table 2.

Baseline and 1 Week performances for BVMT-R and HVLTR learning and learning slope measures – along with PE scores, *p* values, and effect sizes – in Cognitively Normal participants (*n* = 152)

	Baseline Mean (SD)	1 Week Mean (SD)	PE Score Mean (SD)	<i>p</i> Value	<i>Cohen's d</i>
<i>BVMT-R</i>					
Total Recall	18.20 (6.1)	27.53 (5.9)	9.32 (4.7)	<0.001	1.98
Trial One	3.57 (1.9)	8.24 (2.5)	4.66 (2.4)	<0.001	1.93
Trial Two	6.53 (2.4)	9.35 (2.1)	2.82 (1.9)	<0.001	1.46
Trial Three	8.11 (2.3)	9.97 (1.9)	1.85 (1.7)	<0.001	1.07
LR	0.56 (0.2)	0.53 (0.4)	-0.032 (0.4)	0.31	-0.08
<i>HVLTR</i>					
Total Recall	27.42 (4.5)	31.28 (4.0)	3.86 (3.5)	<0.001	1.12
Trial One	7.33 (1.8)	9.61 (1.9)	2.28 (1.8)	<0.001	1.29
Trial Two	9.61 (1.8)	10.65 (1.5)	1.05 (1.6)	<0.001	0.66
Trial Three	10.51 (1.6)	11.01 (1.8)	0.49 (1.4)	<0.001	0.35
LR	0.73 (0.3)	0.73 (0.3)	-0.006 (0.4)	0.84	-0.02

Note: BVMT-R = Brief Visual Memory Test – Revised, HVLTR = Hopkins Verbal Learning Test – Revised, PE = Practice Effect (calculated as 1 Week – Baseline), LR = Learning Ratio. *P* value reflects significance of *one-sample t tests* examining whether PE change scores differed from expectation (PE = 0), and *Cohen's d* reflects the effect size of the *one-sample t tests*.

Table 3.

Baseline and 1 Week performances for BVMT-R and HVLTR learning and learning slope measures – along with PE scores, *p* values, and effect sizes – in participants with Mild Cognitive Impairment (*n* = 131).

	Baseline Mean (SD)	1 Week Mean (SD)	PE Score Mean (SD)	<i>p</i> Value	Cohen's <i>d</i>
<i>BVMT-R</i>					
Total Recall	10.95 (5.5)	18.35 (7.8)	7.40 (5.1)	<0.001	1.45
Trial One	1.99 (1.6)	4.95 (2.7)	2.92 (2.2)	<0.001	1.34
Trial Two	3.82 (2.2)	6.25 (2.7)	2.34 (2.1)	<0.001	1.10
Trial Three	4.93 (2.4)	7.10 (2.8)	2.10 (2.3)	<0.001	0.91
LR	0.32 (0.2)	0.36 (0.3)	0.029 (0.3)	0.28	0.10
<i>HVLTR</i>					
Total Recall	22.16 (5.4)	26.53 (5.8)	4.37 (3.5)	<0.001	1.24
Trial One	5.83 (1.9)	7.82 (2.3)	1.98 (1.9)	<0.001	1.05
Trial Two	7.66 (2.1)	9.14 (2.0)	1.48 (1.6)	<0.001	0.93
Trial Three	8.67 (2.1)	9.57 (2.0)	0.90 (1.6)	<0.001	0.57
LR	0.51 (0.3)	0.49 (0.4)	-0.019 (0.4)	0.57	-0.05

Note: BVMT-R = Brief Visual Memory Test – Revised, HVLTR = Hopkins Verbal Learning Test – Revised, PE = Practice Effect (calculated as 1 Week – Baseline), LR = Learning Ratio. *P* value reflects significance of *one-sample t tests* examining whether PE change scores differed from expectation (PE = 0), and *Cohen's d* reflects the effect size of the *one-sample t tests*.