Published in final edited form as:

Clin Neuropsychol. 2022 April ; 36(3): 571–583. doi:10.1080/13854046.2020.1781933.

Practice effects in Mild Cognitive Impairment: A validation of Calamia et al. (2012)

Kevin Duff, **Dustin B. Hammers**

Center for Alzheimer's Care, Imaging and Research, Department of Neurology, University of Utah, Salt Lake City, UT, USA

Abstract

Objective: In a meta-analysis examining practice effects on repeated neuropsychological testing, Calamia et al. (2012) provided information to predict practice effects in healthy and clinical samples across a range of cognitive domains. However, these estimates have not been validated.

Method: The current study used these prediction estimate calculations to predict follow-up scores across one year on a brief battery of neuropsychological tests in a sample of 93 older adults with amnestic Mild Cognitive Impairment. The predicted follow-up scores were compared to observed follow-up scores.

Results: Using Calamia et al. model's intercept, age, retest interval, clinical status, and specific cognitive tests, 3 of the 7 observed follow-up scores in this cognitive battery were significantly lower than the Calamia et al. predicted follow-up scores. Differences between individual participants' observed and predicted follow-up scores were more striking. For example, on Delayed Recall of the Hopkins Verbal Learning Test - Revised, 40% of the sample had Calamia et al. predicted scores that were one or more standard deviations above their observed scores. These differences were most notable on tests that were not in Calamia et al.'s cognitive battery, suggesting the meta-analysis results may not generalize as well to other tests.

Conclusions: Although Calamia et al. provided a method for predicting practice effects and follow-up scores, these results raise caution when using them in MCI, especially on cognitive tests that were not in their meta-analysis.

Keywords

Practice effects; Mild Cognitive Impairment

Introduction

Practice effects, which are improvements in cognitive test scores due to repeated exposure to testing materials (McCaffrey, Duff, & Westervelt, 2000), are ubiquitous in neuropsychological assessment, both in clinical and research settings. For example, repeat testing and resulting practice effects can occur when tracking progression of a

Address correspondences to Kevin Duff, PhD, Center for Alzheimer's Care, Imaging and Research, Department of Neurology, University of Utah, 650 Komas Drive #106-A, Salt Lake City, UT 84108. Tel: 801-585-9983. kevin.duff@hsc.utah.edu.

disease, monitoring recovery due to an intervention, or re-evaluating a claimant in a forensic case. These “artificial” boosts in cognitive test scores seem to be moderated by multiple factors, including the age, education, retest interval, cognitive domain, and clinical condition (Basso, Carona, Lowery, & Axelrod, 2002; Calamia, Markon, & Tranel, 2012; McCaffrey & Westervelt, 1995; Rapport, Brines, Axelrod, & Theisen, 1997; Salthouse, 2010). Multiple methods have been devised to quantify, control, or mitigate practice effects in neuropsychology (e.g., alternate test forms, dual baseline research design, standardized regression-based change scores) (Beglinger et al., 2005; K. Duff, 2012; K. Duff, Westervelt, McCaffrey, & Haase, 2001; McSweeny, Naugle, Chelune, & Luders, 1993), as practice effects can interfere with the interpretation of longitudinal studies with repeat cognitive testing (Goldberg, Harvey, Wesnes, Synder, & Schneider, 2015; Rabbitt, Diggle, Smith, Holland, & Mc Innes, 2001).

These artificial improvements in test scores may be particularly salient in late-life cognitive disorders. Cooper et al. (Cooper et al., 2001; Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004) showed that patients with Alzheimer’s disease and Mild Cognitive Impairment (MCI) failed to show practice effects on a semantic fluency test that was repeatedly administered to them. However, more recent studies have noted that practice effects may provide useful information about diagnosis, prognosis, and treatment response in MCI and preclinical Alzheimer’s disease (K Duff et al., 2007; K. Duff, Beglinger, Moser, Schultz, & Paulsen, 2010; K. Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2017; Machulda et al., 2014). Furthermore, the absence of practice effects is also associated with smaller hippocampi and increased brain amyloid burden in patients with MCI (K. Duff et al., 2018; K. Duff, Foster, & Hoffman, 2014). Finally, since patients with MCI are at risk for conversion to dementia, they are likely to be seen for repeat neuropsychological assessment.

Calamia et al. (2012) conducted a meta-analysis of practice effects in 379 studies, in which certain, widely-used neuropsychological tests were administered on at least two occasions to adults (healthy controls or clinical patients). Relevant data extracted from these studies included: retest interval, age of participants, sample type, use of a placebo, and use of an alternate form. Scores for tests were presented individually and collapsed into cognitive domains. The age of the participants in these studies were largely 40 – 50 years old, and the retest interval was largely one year. A baseline model was created that identified the expected practice effect of a healthy 40-year old person retested with the same test form of an auditory attention/working memory test after one year. In this baseline model, the average practice effect was 0.242, or about a quarter of a standard deviation unit. Table 2 of that paper displayed beta weights to fit this baseline model to individuals who differed on relevant variables. For example, if an alternate form was used on retesting, 0.217 was subtracted from the baseline model to predict a much smaller practice effect (0.242 – 0.217 = 0.025, 1/40 of a standard deviation unit). Adjustments for other variables (e.g., age, retest interval, clinical condition, cognitive domain assessed) can also be made to the model depending on each unique situation. Table 3 of that paper included beta weights for individual neuropsychological tests. When the practice effects and other adjustments were added to/subtracted from the baseline score, then it would predict an individual’s follow-up score.

Although the results from Calamia et al.'s meta-analysis have the potential to inform clinicians and researcher about the amount of expected change on retesting in their patients and participants, the results have never been validated against an independent sample. Validation of the results of the meta-analysis would provide the field with a greater understanding of how much practice effects vary. Additionally, since the current paper sought to calculate Calamia et al. predicted practice effects/follow-up scores and compare them to observed practice effects/follow-up scores in a sample of patients with amnestic MCI, the current findings might indicate how the predicted scores of Calamia et al. would generalize to an independent, older, and more impaired sample, which is arguably more representative of where such methods will be used in research and clinical settings. Finally, the current study used some cognitive measures that were exact matches to those in the meta-analysis, whereas other measures were different but from the same cognitive domains. It was hoped that using this range of cognitive measures would inform the field as to how much the results of Calamia et al. could be "stretched" in their application to other new samples and settings. Overall, it was hypothesized that the Calamia et al. model would approximate observed test-retest change in this sample. However, the extant literature on practice effects suggests that they are influenced by many variables (e.g., age, education, retest interval, cognitive domains assessed), such that different individuals may show very different amounts of improvement on repeat testing. As such, it is reasonable to assume that group differences in practice effects/follow-up scores may contain a lot of individual variability. Therefore, it is also hypothesized that there would be sizeable numbers of participants whose Calamia et al. predicted practice effects/follow-up scores would not match their observed follow-up scores.

Methods

Participants

Ninety-three older adults (age: M = 74.9 years, SD = 6.0; education: M = 16.5 years, SD = 2.8; 56% male; 98% Caucasian; Reading subtest on the Wide Range Achievement Test – 4: M = 108.6 standard score points, SD = 9.2; 30-item Geriatric Depression Scale: M = 3.8, SD = 3.0) diagnosed with MCI (21% single domain amnestic, 79% multidomain amnestic) participated in a study on memory and aging. They were primarily recruited from a memory disorders clinic, and diagnostic criteria included concern about cognitive change, objective impairment in memory, and preservation of independence of daily activities (Albert et al., 2011). No biomarkers were considered in the classification. Participants provided informed consent before proceeding with the study, and they were compensated for their participation.

Procedures

As part of the study, all participants completed a brief neuropsychological battery at baseline and after approximately 1.3 years (SD = 0.1). The battery included the following tests:

- Trail Making Test Parts A and B (TMT-A, TMT-B; Reitan, 1992) are tests of visual scanning/processing speed and set shifting, respectively. For each part, the score is the time to complete the task. Normative data reverses the score,

so higher values will indicate better performance. These two tests exactly match those in Calamia et al.

- Symbol Digit Modalities Test (SDMT; Smith, 1973) is a divided attention and psychomotor speed task, with the number of correct symbol-digit pairings in 90 seconds being the total score (range = 0 – 110). This test is nearly identical to one of the tests in Calamia et al. (Wechsler Adult Intelligence Scale Digit Symbol Coding).
- Hopkins Verbal Learning Test - Revised (HVLT-R; Brandt & Benedict, 2001) is a verbal learning task of 12 words over three learning trials, with correct words summed for the Total Recall score (range = 0 – 36). The Delayed Recall score is the number of correct words recalled after a 20 – 25 minute delay (range = 0 – 12). This test is similar to one of the tests in Calamia et al.'s Verbal Memory domain (California Verbal Learning Test).
- Brief Visuospatial Memory Test - Revised (BVMT-R; Benedict, 1997) is a visual learning task of six geometric designs in six locations on a card over three learning trials, with correct designs and locations summed for the Total Recall score (range = 0 – 36). The Delayed Recall score is the number of correct designs and locations recalled after a 20 – 25 minute delay (range = 0 – 12). This test is different from the other tests in the Visual Memory domain of Calamia et al. (e.g. Benton Visual Retention Test, Complex Figure Test).

Raw scores were converted into T-scores (M = 50, SD = 10) based on the test manual (HVLT-R, BVMT-R) or existing normative data for older adults (SDMT, TMT-A, TMT-B) (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996; Ivnik et al., 1992).

Statistical Analyses

Using Table 2 in Calamia et al. (2012), predicted practice effects values were calculated for each participant for the seven scores within the brief battery. Each prediction equation is detailed in Table 1. The resulting value from each equation was multiplied by 10 (i.e., the standard deviation of T-scores). The resulting values were then added to the participant's baseline score to yield a "Calamia et al. predicted follow-up score".

Calamia et al. predicted follow-up scores were compared to their respective observed follow-up scores in three ways. First, to compare these two sets of scores (observed and Calamia et al. predicted) on a group level, a series of seven dependent t-tests were calculated. Such results may be helpful in determining the value of Calamia et al.'s model in research, where one wants to know if a group has shown the expected practice effect.

Second, to examine the value of these Calamia et al. predicted scores in individuals, which may be more relevant in clinical settings, z-scores were calculated on two sets of scores using the means and standard deviations of the observed follow-up scores for each of the seven cognitive scores (see Table 1). The observed z-scores, when standardized by themselves, would reflect a relatively normal distribution, and these observed z-scores would serve as the comparator to the Calamia et al. predicted z-scores. The Calamia et al. predicted z-scores, when compared to the observed z-scores, would serve as a test of the hypothesis

that these z-scores were more variable than what was actually observed on follow-up after approximately one year. The two sets of z-scores were quintisected, or divided into five groups: z ≤ −2.00 was coded as −2, z = −1.99 thru −1.00 was coded as −1, z = −0.99 thru 0.99 was coded as 0, z = 1.00 thru 1.99 was coded as 1, and z ≥ 2.00 was coded as 2. This coding scheme is similar to the standard deviations, in z-score units, on a normal distribution curve, such that −2 this reflects the lowest 2.5% of the z-scores, −1 reflects the next 13.5% of z-scores, 0 reflects the middle 68% of z-scores, 1 reflects the next 13.5% of z-scores, and 2 reflects the highest 2.5% of z-scores. The quintisected z-scores for the two sets (observed and Calamia et al. predicted) were compared using chi-square analyses for each cognitive score. In this chi-square analysis, “matches” were defined as the observed and Calamia et al. predicted scores falling within the same standard deviation unit (e.g., observed = −1, Calamia et al. predicted = −1) and “mismatches” were defined as the observed score being greater or less than the Calamia et al. predicted score by at least one standard deviation unit (e.g., observed = 1, Calamia et al. predicted = 0; observed = 0, Calamia et al. predicted = 1). As there were fourteen primary analyses (seven dependent t-tests and seven chi-squares), an alpha value of 0.05 was used throughout.

Third, to get an additional metric of individual differences between observed and Calamia et al. predicted follow-up scores, Lin’s Concordance Correlation Coefficients (CCC) and Root Mean Square Differences (RMSD) (Barchard, 2012; Lin, 1989) were calculated for the quintisected z-scores for each of the seven cognitive scores. Whereas traditional measures of agreement (e.g., Pearson correlation, intraclass correlation coefficient) ignore differences in means and standard deviations between two sets of data, CCC and RMD quantify these differences (Barchard, 2012). Much like a traditional measure of agreement, CCC ranges from −1 to 1. The RMSD ranges from 0 (complete agreement of scores in the two sets) to the difference between the highest and lowest possible values in the two sets, which in the current data would be 3 or 4, depending on the cognitive test.

Results

Baseline, observed follow-up, and Calamia et al. predicted follow-up T-scores for the seven scores from the brief battery in our current sample are presented in Table 2. Consistent with a diagnosis of amnestic MCI, the mean scores on the HVLT-R and BVMT-R were well below expectations, but the scores on the three non-memory tests were closer to the population mean.

Dependent t-tests comparing observed follow-up and predicted follow-up scores were statistically significantly different for 3 of the 7 cognitive scores: HVLT-R Delayed Recall: $t(92)=5.82$, $p<0.001$, $d=0.63$; Brief Visuospatial Memory Test – Revised Total Recall: $t(92)=5.04$, $p<0.001$, $d=0.53$; and Brief Visuospatial Memory Test – Revised Delayed Recall: $t(92)=3.77$, $p<0.001$, $d=0.40$. For each significant difference, the observed follow-up score was significantly lower than the predicted follow-up score.

Using the means and standard deviations for observed follow-up scores for the seven cognitive scores, z-scores were calculated for both the observed follow-up scores and the Calamia et al. predicted follow-up scores. The mean z-score for the observed follow-up

scores ranged from −0.003 to 0.003, which is not surprising since they were standardized against themselves. The mean z-score for the Calamia et al. predicted follow-up scores were larger, ranging from −0.019 to 0.507. The z-scores for each individual participant was quintisected as: −2, −1, 0, 1, or 2. All seven chi-squares that compared the observed and Calamia et al. predicted follow-up values were statistically significant: TMT-A: $\chi^2(16)=83.7$, p<0.001; TMT-B: $\chi^2(16)=132.7$, p<0.001; SDMT: $\chi^2(16)=54.9$, p<0.001; HVLT-R Total Recall: $\chi^2(9)=57.7$, p<0.001; HVLT-R Delayed Recall: $\chi^2(6)=22.5$, p=0.001; BVMT-R Total Recall: $\chi^2(9)=23.8$, p<0.001; and BVMT-R Delayed Recall: $\chi^2(6)=11.2$, p=0.025. For the BVMT-R Total and Delayed Recall, HVLT-R Delayed Recall, and SDMT, the observed quintisected z-scores tended to be one or more standard deviations smaller than those predicted by Calamia et al. Table 3 presents the percentage of “matches” (i.e., observed and Calamia et al. predicted scores falling within the same standard deviation unit) and “mismatches” (i.e., observed score being greater or less than the Calamia et al. predicted score by at least one standard deviation unit) for each of the seven cognitive test scores.

Table 3 also presents the CCC and RMSD for each of the seven cognitive test scores. The CCC values tend to indicate poor agreement between the observed and the Calamia et al. predicted follow-up scores (e.g., 0.15 – 0.71). Similarly, the RMSD values largely approached one (0.55 – 0.81), which indicates differences between the two sets of follow-up scores.

Discussion

In a meta-analysis, Calamia et al. (2012) quantified the amount of expected practice effect on widely-used neuropsychological tests administered twice to healthy controls or clinical patients. In a baseline model (i.e., a healthy 40-year old person retested with the same test form of an auditory attention/working memory test after one year), they found that the average practice effect was about a quarter of a standard deviation unit. Furthermore, they provided adjustments that would increase or decrease this predicted practice effect depending on the age and clinical condition of the individual, retest interval, or cognitive domain being assessed. Although such a model can be useful in predicting practice effects and follow-up scores in clinical and research settings, this model has never been validated with an independent sample. Therefore, in 93 older participants with amnestic MCI, the values in Calamia et al. were used to predict practice effects and follow-up scores, and those Calamia et al. predicted follow-up scores were compared to observed follow-up scores in these same participants. Results indicated that the mean predicted follow-up scores using the Calamia et al. model were significantly higher than the observed follow-up scores for 3 of the 7 neuropsychological test scores (all memory scores). When individual observed and Calamia et al. predicted follow-up scores were compared, more striking differences emerged. There were considerable differences for individual participants, with mismatches being quite prevalent, especially for the memory measures. For some neuropsychological test scores, Calamia et al. predicted scores that were one or more standard deviations greater than observed scores, and other test scores showing more comparable observed and predicted follow-up scores.

Overall, these results highlight the complexity of assessing change across time, including practice effects, tracking disease progression, and response to an intervention. On a group level, the prediction estimates from Calamia et al. (2012) performed quite well on measures of processing speed and executive functioning. For example, there were minimal differences between these observed and predicted follow-up scores (mean differences 0.1 – 1.4 T-score points). However, on tests of learning and memory, the current study raises concern about these prediction estimates. For example, on these tests, the differences were much larger, with mean differences between observed and predicted follow-up scores ranging from 1 – 7 T-score points. As such, it appears that the model of Calamia et al. should be used with caution on a group level when predicting memory scores. If an individual was conducting a pilot research study, did not have sufficient resources for a control group, and was planning to use Calamia et al.'s model to predict the amount of change that was considered "normal" or "typical," then this researcher should consider that the predicted memory scores might be higher than what would be found if actual retesting occurred in a control group.

When individual cases were examined, more consistent concern surfaces about the applicability of this prediction model. In the current sample, although many matches occurred between the observed follow-up scores and the Calamia et al. predicted scores (e.g., observed = 0, Calamia et al. predicted = 0; see Table 3), there were also many mismatches in these two sets of scores. For example, on the Delayed Recall trial of the BVMT-R (see Table 3), only 48% of Calamia et al. predicted scores fell within the same standard deviation (e.g., both 0), but 37% of the Calamia et al. predicted z-scores were categorized as higher than the observed scores (e.g., observed = −1, Calamia et al. predicted = 0), and approximately 15% of Calamia et al. predicted z-scores were less than the observed scores (e.g., observed = 1, Calamia et al. predicted = 0). Compared to the expectations of a normal distribution, a significantly greater number of mismatches, by one or more standard deviation categories, were seen on all seven cognitive test scores. These mismatches were further highlighted by Lin's CCC and RMSD. The CCC (which can be interpreted similar to other correlations) ranged from 0.15 – 0.71, and this appears well below ideal expectations for clinical activity. The RMSD, which would yield values of near-0 if there was a lot of concordance between the observed and predicted follow-up scores, had values that also suggested poor agreement among the two sets of the seven cognitive test scores. As such, it appears that the model of Calamia et al. may be less applicable to individual cases. Consequently, if an individual was assessing cognitive change in a clinical patient, the Calamia et al. model might predict an accurate amount of change (i.e., match), or it might over- or under-predict the expected practice effect and follow-up score by a significant amount (i.e., mismatch).

It should be reiterated that the current study included a range of neuropsychological measures, some of which were included in Calamia et al.'s meta-analysis and others that were not. As such, a secondary purpose of this study was to examine the degree to which measures that "stretched" from Calamia et al.'s cognitive battery could still be predicted from the results of their meta-analysis. Exact matches of tests included TMT-A and TMT-B, with SDMT being a very near match. HVLT-R was considered a minor "stretch," as it is quite similar to the California Verbal Learning Test from Calamia et al.'s model. Finally, the BVMT-R was considered a major "stretch," as it was quite different from

the Complex Figure Test and Benton Visual Retention Test but it was still in the same cognitive domain of visual memory. For the three exact/near-exact matches, none showed differences between the observed and predicted follow-up scores. On individual analyses, all three showed relatively high levels of matches (TMT-A=71%, TMT-B=81%, SDMT=73%). Therefore, in cases with exact/near-exact matches of the cognitive tests, the Calamia et al.'s prediction estimate calculations showed adequate accuracy at denoting a follow-up score. Conversely, differences were more apparent on tests that "stretched" from those reported in the meta-analysis. For example, on the HVLT-R (which was a minor "stretch" from the California Verbal Learning Test), modest differences were seen between the observed and predicted follow-up scores for Delayed Recall (d=0.69), with the predicted score being approximately 7 T-score points higher than the observed score. Individual analyses also showed that this cognitive score had one of the highest RMSD, suggesting the lowest agreement between observed and predicted follow-up scores. On the BVMT-R (which was considered a major "stretch" from the Complex Figure Test and Benton Visual Retention Test), also showed medium to large differences between observed and predicted follow-up scores (Total Recall d=0.67, Delayed Recall d=0.51). Since the two "stretch" tests (HVLT-R and BVMT-R) showed the largest differences between observed and predicted follow-up scores, Calamia et al.'s model might not be particularly accurate for tests that were not part of their meta-analysis. Even though the predicted scores were significantly different from the observed scores on the tests with more "stretch," there remains a need to "stretch" our change formulae, as it unusual to find exact matches of neuropsychological measures when a patient is tested twice in a clinical setting.

To our knowledge, this is the first attempt to validate the results of the meta-analysis by Calamia et al. (2012). Lubrini et al. (2020) used these findings to estimate the amount of practice effects in a small sample of patients with traumatic brain injuries, who were evaluated shortly after their injuries and six months later. By estimating the expected practice effects in their sample with the model of Calamia et al., they were then able to quantify how much "normal" recovery was occurring. For example, they estimated that 36% of the change seen on the TMT-B was due to practice effect, and the remaining 64% of the total change was due to normal recovery. Although this seems to be an appropriate use of the data in Calamia et al., there is some concern about the amount of mismatch of predicted and observed follow-up scores in Lubrini et al.'s small cohort. As such, it would be difficult to determine, with much precision, if the cognitive change variance attributed to practice effects and recovery are accurate.

There may be some reasons for the significant group differences and the prevalent mismatches that were seen in the individual comparisons. First, the current sample consisted of older adults diagnosed with amnestic MCI. Such a group may show more variable practice effects than the other clinical samples in Calamia et al. (2012). For example, we have shown that there appear to be two subgroups in such a sample: those that show a traditional practice effect on repeat testing and those that show a much smaller practice effect (K. Duff et al., 2008). Additionally, these differential rate of practice seem to portend future cognitive functioning (K. Duff et al., 2011). Even though Calamia et al. had studies with MCI in their meta-analysis, their effect in the overall finding may have been diminished because so many other clinical conditions were also included. Second, our sample was

mostly multidomain MCI, which may be closer to frank Alzheimer's disease than cases of single domain. In Calamia et al., it is not clear if their MCI category is referring to the single or multidomain subtype. However, it is clear that the beta weight for Alzheimer's disease is nearly nine times higher than the beta weight for MCI (−0.517 vs. −0.058, respectively). As such, it is reasonable to suspect that the studies in the meta-analysis may have fallen on the milder end of this disease severity spectrum than those in the current study. Third, as detailed earlier, the current study used neuropsychological tests that only partially aligned with those reported by Calamia et al. Although some tests were exact matches, others "stretched" from measures in the meta-analysis. In our study, the accuracy of the prediction estimate calculations were slightly better in the exact tests reviewed by Calamia et al compared to our "stretch" tests. Fourth, there may be other variables that seem to impact practice effect that were not included in Calamia et al.'s model. For example, education and intelligence are two factors that are positively related to practice effects that were not part of this meta-analysis. This may be relevant in the current study because our sample of patients with MCI were highly educated (mean = 16.5 years, range = 12 – 22 years). Relatedly, it is possible that differences were present because our cohort was notably older than the average participants in the meta-analysis of Calamia et al. (70's vs. 40–50's, respectively), and the age estimation of Calamia et al. (i.e., −0.004 for every year over 40) may need some adjustment to improve its prediction accuracy in older individuals. These differences between the current study and the meta-analysis of Calamia et al. are not necessarily weaknesses of either, but they may explain some of the discrepant results.

As noted in the beginning of this paper, repeat testing and practice effects are ubiquitous in neuropsychological assessment, both in clinical and research settings. Whereas Calamia et al. (2012) provide one method of addressing this "problem," there are other ways to address practice effects. For example, there are alterations in the design of a study or patient interaction that can minimize these artificial improvements in test scores on follow-up testing. The use of a well-matched control group, alternate test forms, or a dual baseline approach can either reduce or accurately model practice effects to inform the clinician or researcher. Similarly, there are a host of statistical procedures that can be used to address practice effects (K. Duff, 2012). Reliable Change Indexes, preferably those that control for practice effects, or Standardized Regression-Based change formulae are two such approaches that have been widely used in neuropsychology. As also noted earlier, the assessment of cognitive change is complex, and it may require that a researcher or clinician utilize multiple methods to most accurately determine how much of a follow-up score is real or artifact.

The current study is not without its limitations. As mentioned previously, we used a brief, research battery of neuropsychological tests that covered the domains of memory and processing speed. The accuracy of Calamia et al.'s prediction calculations may have been higher in other cognitive tests. The current study only tested a few of the variables in the meta-analysis (e.g., baseline practice effect, age, retest interval, a single clinical condition, cognitive domain). Other variables (e.g., use of an alternate form, use of a placebo, other clinical conditions) were not present in our current study, and they could not be examined. A wider test of the baseline model of Calamia et al. is needed to see if it fares better in other situations. Since many participants in the current study were recruited from a memory

disorders clinic, it is possible (perhaps likely) that they had been previously exposed to some/all of the current neuropsychological measures. It is unclear how prior exposure to these tests may have influenced their current performances, which could influence how those performances related to predicted scores. Finally, although there are multiple methods for assisting clinicians and researchers in determining if a reliable change in cognition has occurred across time, including alternate test forms, dual baseline research design, and standardized regression-based change scores, the current study cannot indicate if any of these methods are better or worse than the method offered by Calamia et al. Despite these limitations, the current study was an initial attempt to validate the impressive work of Calamia et al. Although the current results only partially validated their efforts in MCI, future work is needed in this area to more comprehensively assist clinicians and researchers in understanding cognitive change.

Acknowledgements:

The project described was supported by research grants from the National Institutes on Aging: R01AG045163 and R01AG055428. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

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. Alzheimers Dement, 7(3), 270–279. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21514249 [PubMed: 21514249]

Barchard KA (2012). Examining the reliability of interval level data using root mean square differences and concordance correlation coefficients. Psychol Methods, 17(2), 294–308. doi:10.1037/a0023351 [PubMed: 21574711]

Basso MR, Carona FD, Lowery N, & Axelrod BN (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. Clin Neuropsychol, 16(1), 57–63. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11992227 [PubMed: 11992227]

Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, … Siemers ER (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol, 20(4), 517–529. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15896564 [PubMed: 15896564]

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. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22540222 [PubMed: 22540222]

Cooper DB, Epker M, Lacritz L, Weine M, Rosenberg RN, Honig L, & Cullum CM (2001). Effects of practice on category fluency in Alzheimer's disease. Clin Neuropsychol, 15(1), 125–128. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11778573 [PubMed: 11778573]

Cooper DB, Lacritz LH, Weiner MF, Rosenberg RN, & Cullum CM (2004). Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. Alzheimer Dis Assoc Disord, 18(3), 120–122. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15494616 [PubMed: 15494616]

Duff K (2012). Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Arch Clin

Neuropsychol, 27(3), 248–261. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22382384 [PubMed: 22382384]

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, Beglinger L, Schultz S, Moser D, McCaffrey R, Haase R, … Paulsen J (2007). Practice effects in the prediction of long-term cognitive outcome in three patient samples: A novel prognostic index. Arch Clin Neuropsychol, 22(1), 15–24. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17142007 [PubMed: 17142007]

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 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20104658 [PubMed: 20104658]

Duff K, Beglinger LJ, Van Der Heiden S, Moser DJ, Arndt S, Schultz SK, & Paulsen JS (2008). Short-term practice effects in amnestic mild cognitive impairment: implications for diagnosis and treatment. Int Psychogeriatr, 20(5), 986–999. doi:S1041610208007254 [pii] 10.1017/S1041610208007254 [PubMed: 18405398]

Duff K, Foster NL, & Hoffman JM (2014). Practice effects and amyloid deposition: preliminary data on a method for enriching samples in clinical trials. Alzheimer Dis Assoc Disord, 28(3), 247–252. doi:10.1097/WAD.0000000000000021 [PubMed: 24614265]

Duff K, Lyketsos CG, Beglinger LJ, Chelune G, Moser DJ, Arndt S, … McCaffrey RJ (2011). Practice effects predict cognitive outcome in amnestic mild cognitive impairment. Am J Geriatr Psychiatry, 19(11), 932–939. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=22024617 [PubMed: 22024617]

Duff K, Westervelt HJ, McCaffrey RJ, & Haase RF (2001). Practice effects, test-retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. Arch Clin Neuropsychol, 16(5), 461–476. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=14590160 [PubMed: 14590160]

Goldberg TE, Harvey PD, Wesnes KA, Synder PJ, & Schneider LS (2015). Practice effects due to serial cognitive assessment: Implications for preclinical Alzheimer's disease randomized controlled trials. Alzheimer's & Dementia, 1(1), 103–111.

Hassenstab J, Ruvolo D, Jasielec M, Xiong C, Grant E, & Morris JC (2015). Absence of practice effects in preclinical Alzheimer's disease. Neuropsychology, 29(6), 940–948. doi:10.1037/neu0000208 [PubMed: 26011114]

Ivnik RJ, Malec JF, Smith GE, Tangalos EG, & Petersen RC (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE Token, WRAT-R Reading, AMNART, STROOP, TMT, and JLO. The Clinical Neuropsychologist, 10, 262–278.

Ivnik RJ, Malec JF, Smith GE, Tangalos EG, Petersen RC, Kokmen E, & Kurland LT (1992). Mayo's Older Americans Normative Studies: WAIS-R norms for ages 56 to 97. The Clinical Neuropsychologist, 6, 1–30.

Lin LI (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), 255–268. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2720055 [PubMed: 2720055]

Lubrini G, Viejo-Sobera R, Perianez JA, Cicuendez M, Castano AM, Gonzalez-Marques J, … Rios-Lago M (2020). Evolution of cognitive impairment after a traumatic brain injury: is there any improvement after controlling the practice effect? Rev Neurol, 70(2), 37–44. doi:10.33588/rn.7002.2019233 [PubMed: 31930469]

Machulda MM, Hagen CE, Wiste HJ, Mielke MM, Knopman DS, Roberts RO, … Petersen RC (2017). Practice effects and longitudinal cognitive change in clinically normal older adults differ by Alzheimer imaging biomarker status. Clin Neuropsychol, 31(1), 99–117. doi:10.1080/13854046.2016.1241303 [PubMed: 27724156]

Machulda MM, Pankratz VS, Christianson TJ, Ivnik RJ, Mielke MM, Roberts RO, … Petersen RC (2014). 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(8), 1247–1264. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=24041121

McCaffrey RJ, Duff K, & Westervelt HJ (2000). Practitioner's Guide to Evaluating Change with Neuropsychological Assessment Instruments. New York: Plenum/Kluwer.

McCaffrey RJ, & Westervelt HJ (1995). Issues associated with repeated neuropsychological assessments. Neuropsychol Rev, 5(3), 203–221. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8653109 [PubMed: 8653109]

McSweeny A, Naugle RI, Chelune GJ, & Luders H (1993). "T Scores for Change": An illustration of a regression approach to depicting change in clinical neuropsychology. Clinical Neuropsychologist, 7(3), 300–312.

Rabbitt P, Diggle P, Smith D, Holland F, & 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(5), 532–543. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11254936 [PubMed: 11254936]

Rapport LJ, Brines DB, Axelrod BN, & Theisen ME (1997). Full scale IQ as a mediator of practice effects: The rich get richer. The Clinical Neuropsychologist, 11(4), 375–380.

Salthouse TA (2010). Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology, 24(5), 563–572. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=20804244 [PubMed: 20804244]

Table 1.

Practice effects prediction equations for each cognitive score.

Cognitive score	Calamia et al. predicted practice effect
TMT-A	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (0.007)
TMT-B	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (−0.008)
SDMT	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (0.007)
HVLT-R Total Recall	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (−0.007)
HVLT-R Delayed Recall	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (−0.007)
BVMT-R Total Recall	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (0.099)
BVMT-R Delayed Recall	(0.242) + (age*−0.004) + (retest*−0.058) + (−0.058) + (0.099)

Note. TMT = Trail Making Test, SDMT = Symbol Digit Modalities Test, HVLT-R = Hopkins Verbal Learning Test – Revised, BVMT-R = Brief Visuospatial Memory Test – Revised. In each Calamia et al. predicted score, the first variable is the baseline model's practice effect, the second variable is the participant's age – 40, the third variable is the retest interval in years, the fourth variable is for the type of sample (i.e., Mild Cognitive Impairment), and final variable is for the cognitive test from Calamia et al. (2012). For TMT-A and SDMT, the beta weight for "Processing Speed" was used (0.007). For the TMT-B, the beta weight for "Executive Functioning" was used (−0.008). For the HVLT-R Total and Delayed Recall, the beta weight for "Verbal Memory" was used (−0.007). For the BVMT-R Total and Delayed Recall, the beta weight for "Visual Memory" was used (0.099).

Table 2.

Baseline, observed follow-up, and Calamia et al. predicted follow-up scores for the seven cognitive scores.

Cognitive score	Baseline (n=93)	Observed follow-up (n=93)	Calamia et al. Predicted follow-up
TMT-A	50.1 (9.5)	50.5 (9.8)	49.9 (9.5)
TMT-B	47.7 (10.5)	47.5 (12.1)	47.4 (10.6)
SDMT	48.7 (10.4)	47.0 (11.7)	48.4 (10.4)
HVLT-R Total Recall	40.3 (9.7)	40.5 (11.2)	39.9 (9.6)
HVLT-R Delayed Recall	33.1 (12.8)	33.0 (14.0)	39.9 (9.6)
BVMT-R Total Recall	34.3 (9.5)	35.0 (11.8)	41.0 (9.6)
BVMT-R Delayed Recall	34.5 (11.6)	35.8 (14.2)	41.0 (9.6)

Note. TMT = Trail Making Test, SDMT = Symbol Digit Modalities Test, HVLT-R = Hopkins Verbal Learning Test – Revised, BVMT-R = Brief Visuospatial Memory Test – Revised. All scores are T-scores (M = 50, SD = 10). Predicted follow-up scores come from application of Table 2 in Calamia et al. (2012).

Table 3.

Matches and mismatches and CCC and RMSD for the observed and Calamia et al. predicted scores for the seven cognitive test scores.

Cognitive score	Matches	Mismatch (O>P)	Mismatch (P>O)	CCC	RMSD
TMT-A	71%	16%	12%	0.55	0.62
TMT-B	81%	10%	9%	0.71	0.55
SDMT	73%	10%	17%	0.41	0.70
HVLT-R Total Recall	70%	17%	13%	0.53	0.59
HVLT-R Delayed Recall	48%	12%	40%	0.32	0.79
BVMT-R Total Recall	54%	14%	32%	0.28	0.79
BVMT-R Delayed Recall	48%	15%	37%	0.15	0.81

Note. TMT = Trail Making Test, SDMT = Symbol Digit Modalities Test, HVLT-R = Hopkins Verbal Learning Test – Revised, BVMT-R = Brief Visuospatial Memory Test – Revised, CCC = Lin's Concordance Correlation Coefficient, RMSD = Root Mean Square Difference, O>P = Observed score was greater than Predicted score, P>O = Predicted score was greater than Observed score.