HHS Public Access
Author manuscript
Neuropsychol Dev Cogn B Aging Neuropsychol Cogn. Author manuscript; available in PMC 2022 May 01.

Published in final edited form as:
Neuropsychol Dev Cogn B Aging Neuropsychol Cogn. 2021 May ; 28(3): 472–492. doi:10.1080/13825585.2020.1787942.

Validation of one-week reliable change methods in cognitively intact community-dwelling older adults.

Dustin B. Hammers[1,2], **Kayla R. Suhrie**[1], **Ava Dixon**[1], **Sariah Porter**[1], **Kevin Duff**[1,2]

[1]Center for Alzheimer's Care, Imaging, and Research, Department of Neurology, University of Utah

[2]Center on Aging, University of Utah

Abstract

Objective: Reliable change methods can assist determination of whether observed changes in performance are meaningful. The current study sought to validate previously-published standardized regression-based (SRB) equations for commonly-administered cognitive tests using a cognitively-intact sample of older adults, and extend findings by including relevant demographic and test-related variables known to predict cognitive performance.

Method: This study applied Duff's (2014) SRB prediction equations to 107 cognitively-intact older adults assessed twice over one week. Prediction equations were also updated by pooling the current validation sample with 93 cognitively-intact participants from Duff's original development sample to create a combined development sample.

Results: Significant improvements were seen between observed baseline and follow-up scores on most measures. However, few differences were seen between observed follow-up scores and those predicted from Duff's SRB algorithms, and the level of practice effects observed based on Duff's equations were consistent with expectations. When SRBs were re-calculated from this combined development sample, predicted follow-up scores were mostly comparable with Duff's equations, but standard errors of the estimate were consistently smaller.

Conclusions: These results help support validity of Duff's (2014) SRB equations to predict cognitive performance on these measures when repeated administration is necessary over short intervals. Findings also highlight the utility of expanding SRB models when predicting follow-up performance serially to provide more accurate assessment of reliable change at the level of the individual. As short-term practice effects are shown to predict cognitive performance annually, they possess the potential to inform clinical decision making about individuals along the Alzheimer's continuum.

Keywords

Reliable Change; Assessment; Neuropsychology

Address correspondences to Dustin B. Hammers, 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: 801-585-3929. dustin.hammers@hsc.utah.edu.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

INTRODUCTION

Reliable change methods are statistical procedures developed to distinguish meaningful change in longitudinal or serial neuropsychological assessment from repeated test exposure benefits (i.e., practice effects; Hammers, Duff, & Chelune, 2015; Lezak, Howieson, Bigler, & Tranel, 2012). Of the multiple procedures available, McSweeny and colleagues' (McSweeny, Naugle, Chelune, & Luders, 1993) standardized regression-based (SRB) predicted difference method possesses broad acceptance (Attix et al., 2009; Crockford et al., 2018; Duff, Beglinger, Moser, Paulsen, et al., 2010; Duff et al., 2004; Duff et al., 2005; Gavett, Ashendorf, & Gurnani, 2015; Rinehardt et al., 2010; Sanchez-Benavides et al., 2016; Stein, Luppa, Brahler, Konig, & Riedel-Heller, 2010). Briefly, the complex-SRB method uses linear regression to predict retest scores (Time 2) for individuals based on their baseline (Time 1) performance and other relevant information (e.g., demographics, test-retest interval, etc.), whereas the simple-SRB method uses Time 1 performance as the sole predictor. A discrepancy change score, or z score, is calculated by comparing an individual's predicted and observed Time 2 scores and dividing by the standard error of the estimate (SE_{est}) of the regression model ($z = (T_2 - \hat{T_2}) / SE_{est}$). Discrepancy change scores (z scores) below −1.645 frequently represent "decline" when using reliable change methods, whereas z scores > 1.645 reflect "improvement" and z scores between +/− 1.645 indicate stability. These z score cut-offs are based on the use of 90% confidence intervals (CIs) of stability (McSweeny et al., 1993), such that a z of 1.645 equates to significance at an α value of $\alpha = .10$. Consequently, if the z scores were normally distributed, then one would expect 5% of participants to show "decline," 90% would remain "stable," and 5% would "improve" beyond expectation. When examining change over shorter periods of time in populations not expected to display acute changes in cognition, however, follow-up scores may be better than baseline scores, yet still worse than predictions based on the SRB algorithms. In this case, we suggest that the term "decline" be replaced with "smaller-than-expected practice effects" when z scores are < −1.645. Conversely, an individual could show "larger-than-expected practice effects" (z > 1.645) or "typical practice effects" (z scores between +/− 1.645). For example, suppose an individual was assessed twice on a memory test over one week, and she obtained a raw score of 18 out of 36 at Time 1 and 22 out of 36 at Time 2. This 4-point improvement, in the absence of a reason for this change (e.g., resolution of a medical issue or initiation of a drug known to assist cognition), would be suggestive of a practice effect. However, if an SRB prediction equation for this memory test predicted that her Time 2 performance should have been 27 out of 36 (i.e., a 9-point improvement) – and the difference between her observed and predicted Time 2 scores yielded a z score < −1.645 – she would be displaying both objective improvement between test administrations and significantly smaller-than-expected practice effects at Time 2.

Duff (2014) previously developed regression-based prediction equations for several commonly administered cognitive tests, including the Hopkins Verbal Learning Test – Revised (HVLT-R; Brandt & Benedict, 2001), the Brief Visuospatial Memory Test – Revised (BVMT-R; Benedict, 1997), Symbol Digit Modality Test (SDMT; Smith, 1973), and the Trail Making Test Parts A and B (TMT-A and TMT-B; Reitan, 1992). Each measure was assessed twice over one week in 167 community-dwelling older adults, of whom 93 were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

classified as cognitively intact and 74 were classified as having Mild Cognitive Impairment (MCI). The relatively rapid re-assessment of these measures permits examination of the impact of short-term practice effects on cognitive performance for these commonly used tasks; however, Duff unfortunately never internally validated these SRB equations at the time of publication (Duff, 2014). Subsequent work (Duff et al., 2018; Duff et al., 2017) applied these prediction equations to samples ($n = 25 – 58$) of community-dwelling older adults with varying levels of cognitive abilities, and tended to show that more impaired participants displayed smaller-than-expected practice effects. More recently, in a sample of 143 participants with MCI, incrementally smaller-than-expected benefit from practice was observed for increasingly impaired amnestic MCI subtypes (Hammers, Suhrie, Dixon, Porter, & Duff, 2020).

Despite these more recent findings, no study has attempted to validate Duff’s SRB prediction equations in a clean sample of cognitively intact individuals. This represents a knowledge gap for these SRBs because 44% of Duff’s (2014) development sample were categorized as having MCI, which raises concern about the purity of that sample and subsequently the resultant SRB equations. The use of “clean” or “robust” normative samples is receiving heightened focus in recent years (Goodwill et al., 2019; Harrington et al., 2017), suggesting the importance of normative samples being free from individuals with cognitive impairment. This increased focus is because cognitively impaired participants will display lower-than-expected baseline performances on cognitive tasks and have been shown to possess diminished benefit from practice (Duff et al., 2018; Duff et al., 2017). Inclusion of cognitively impaired participants in SRB development therefore leads to the potential for prediction equations to under-predict Time 2 performance. Consequently, one aim of the current study was to examine the validity of Duff’s (2014) SRB prediction equations using independent samples of cognitively intact community-dwelling older adults. Based on the inclusion of MCI participants in Duff’s sample, it was hypothesized that the application of these prediction equations to intact samples would result in a greater proportion of participants benefiting from practice on these cognitive measures over one week than expected (i.e., under-prediction of equations).

In addition, while Duff’s (2014) SRB equations incorporated demographic factors of age, education, and gender into his published models, additional demographic or test-characteristic factors shown to impact practice effects exist and have the potential to improve prediction of Time 2 scores. For example, research has repeatedly shown an association between premorbid intellect and practice effects (Patton et al., 2005; Rapport et al., 1997; Rapport, Brines, Axelrod, & Theisen, 1997), and it has been suggested that individuals with stronger premorbid intellect or baseline performances on cognitive tasks display greater benefit from repeated exposure to stimuli (Rapport et al., 1997). Also, length of the retest interval between serial assessments has been shown to influence the size of practice effects observed over time (Calamia, Markon, & Tranel, 2012). In a meta-analysis of 349 studies assessing practice effects across a variety of neuropsychological tests, Calamia and colleagues (2012) highlighted retest interval among other patient- (age, diagnostic status) and test-characteristic (use of alternate form or placebo) factors, and observed that shorter retest intervals were associated with increases in estimated score gains at retesting. Relatedly, Duff has shown that one-week retest intervals are particularly susceptible to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

practice effects, and that one-week practice effects add to the prediction of one-year performance on these repeated measures (Duff, Beglinger, Moser, Paulsen, et al., 2010). They observed that while baseline test performance was the strongest predictors of future test performance, one-week practice effects was consistently the second greatest predictor and improved the predictability of one-year repeated performance by 3% to 22%.

Taken together, a second aim of the study was to update the Duff (2014) SRBs by 1) combining the cognitively intact participants from Duff's original development sample with the current validation sample to create a larger and cognitively cleaner combined development sample than Duff's original sample, and 2) adding premorbid intellect and retest interval to the prediction models to examine the impact of these relevant variables on prediction of Time 2 performance during serial assessment. It was hypothesized that increasing the sample size of the combined development sample and including premorbid intellect and retest interval in the models would improve the predictive capacity of these SRB equations for these measures. Further validation of these SRB prediction equations in a large and clean normative sample with a greater set of predictors would increase confidence in these equations when tests are repeated over a short interval, and increase their potential diagnostic and prognostic value for predicting performance over more traditional clinical time-frames.

METHOD

Participants

Cognitively intact community dwelling older adults were recruited from the community (e.g., senior centers and independent living facilities) from two different samples for the current study. The first sample was comprised of 55 cognitively intact community-dwelling older adults recruited from 2010 to 2013 as a control group for a study of practice effects and MCI (see Duff et al., 2017). The second sample was comprised of 52 cognitively intact community-dwelling older adults recruited as a control group for a study of practice effects and Alzheimer's disease biomarkers (2019 to present). The first sample's mean age was 74.1 (*SD* = 6.3, range = 65 – 89) years old, and the second sample was 72.5 (*SD* = 4.9, range = 65 – 91) years old. The first sample averaged 15.4 (*SD* = 2.9, range = 8 – 20+) years of education, with an average of 16.7 (*SD* = 2.1, range = 12 – 20) years of education for the second sample. Both samples were predominantly Caucasian, with the first sample being predominantly female (81.8% female) and the second sample having a slightly higher proportion of females than males (61.5% female). Premorbid intellect at baseline was average according to the Wide Range Achievement Test – fourth edition (WRAT; Wilkinson & Robertson, 2006) Reading subtest for both samples (standard score: *M* = 108.6, *SD* = 8.0, range = 85 – 126 for the first sample, and *M* = 110.6, *SD* = 7.4, range = 88 – 126 for the second sample). Self-reported depression was generally low for both samples, including an average of 4.8 (*SD* = 4.0, range = 0 – 14) according to the 30-item Geriatric Depression Scale (GDS; Yesavage et al., 1982) for the first sample, and an average of 0.9 (*SD* = 1.0, range = 0 – 5) for the second samples using the 15-item GDS (cut-off ≤ 5; Sheikh & Yesavage, 1986; though self-reported depression was part of the exclusionary criteria for the parent study of the latter sample). The mean retest interval was 7.6 (*SD* = 1.9, range = 6 –

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

14) days for the first sample, whereas the retest interval was 6.8 (*SD* = 0.8, range = 4 – 9) days for the second sample.

For inclusion in the study, all participants from both samples were classified as being cognitively intact, or free of cognitive impairment (e.g., MCI or dementia due to Alzheimer's disease). Classification of participants from the first sample has been described previously (Duff et al., 2017). Briefly, all participants in this sample performed within 1.5 SD of the mean for each domain of a baseline cognitive evaluation described below. Classification of participants from the second sample was based on the classification battery developed in the Alzheimer's Disease Neuroimaging Initiative (ADNI2, 2020), which included the Mini Mental Status Examination (Folstein, Folstein, & McHugh, 1975), the Clinical Dementia Rating Scale (Morris, 1993), and the Wechsler Memory Scale-Revised (Wechsler, 1987) Logical Memory II Paragraph A. As can be observed in Table 1, on average the two samples displayed within to above expectation abilities for baseline immediate and delayed memory skills, visuospatial skills, language, and attention on the Repeated Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, 2012).

The two cognitively intact samples differed on retest interval, $t(71.56) = 2.79$, $p = .007$, $d = 0.66$, but there were no significant differences on any other continuous demographic variables (see Table 1). Specifically, no significant differences were observed for variables of age, $t(101.1) = 1.48$, $p = .14$, $d = 0.29$, education, $t(99.0) = -2.58$, $p = .01$, $d = -0.52$, premorbid intellect, $t(105) = -1.34$, $p = .18$, $d = -0.26$, RBANS Immediate Memory Index, $t(105) = 2.51$, $p = .01$, $d = 0.49$, RBANS Visuospatial/Constructional Index, $t(105) = -2.17$, $p = .03$, $d = -0.42$, RBANS Language Index, $t(105) = 0.76$, $p = .45$, $d = 0.15$, RBANS Attention Index, $t(105) = -1.26$, $p = .21$, $d = -0.25$, RBANS Delayed Memory Index, $t(88.18) = -0.04$, $p = .97$, $d = -0.01$, and RBANS Total Scale score, $t(105) = -0.07$, $p = .94$, $d = -0.01$. Similarly, there were no differences on dichotomous demographic variables between the two groups, including for sex, χ^2 (1) = 4.49, $p = .03$, $Phi = 0.23$, or ethnic distribution, χ^2 (1) = 0.95, $p = .33$, $Phi = 0.09$. Overall, any differences in the samples were small in magnitude, therefore these groups were pooled together to create a cognitively intact combined validation sample with a total sample size of 107 participants. Please see Table 1 for the pooled demographic values for the combined validation sample.

Additionally, for the calculation of new SRB equations, the current study pooled the current combined validation sample of 107 cognitively intact participants with the 93 cognitively intact participants from Duff's (2014) development sample, resulting in a total combined development sample of 200 cognitively intact participants. See Table 1 for the demographic characteristics of this combined development sample.

General inclusion criteria for the study involved being aged 65 years or older and functionally independent (according to participant and/or knowledgeable informant), along with possessing adequate vision, hearing, and motor abilities to complete the cognitive evaluation. General exclusion criteria included neurological conditions likely to affect cognition, dementia, major psychiatric condition, current severe depression, substance

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

abuse, anti-convulsant or anti-psychotic medications, or residence in a skilled nursing or living facility.

Procedure

All procedures were approved by the local Institutional Review Board before the study commenced. All participants provided informed consent before completing any procedures. The following measures were administered at a baseline visit:

- HVLT-R (Brandt & Benedict, 2001) is a verbal memory task with 12 words learned over three trials, with the 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). For all HVLT-R scores, higher values indicate better performance.
- BVMT-R (Benedict, 1997) is a visual memory task with 6 geometric designs in 6 locations on a card learned over three 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). For all BVMT-R scores, higher values indicate better performance.
- SDMT (Smith, 1973) is a divided attention and psychomotor speed task, with the number of correct responses in 90 seconds being the total score (range = 0 – 110), and higher values indicate better performance.
- TMT-A and TMT-B (Reitan, 1992) are tests of visual scanning/processing speed and set shifting/complex mental flexibility, respectively. For each part, the score is the time to complete the task (range = 0 – 180 seconds for TMT-A, and range = 0 – 300 seconds for TMT-B). Higher values indicate poorer performance.
- WRAT Reading subtest – fourth edition (Wilkinson & Robertson, 2006) is used as an estimate of premorbid intellect, in which an individual attempts to pronounce irregular words. The score is normalized to standard scores (M = 100, SD = 15) relative to age-matched peers. Higher values indicate better performance.
- RBANS (Randolph, 2012) is a neuropsychological test battery comprising 12 subtests that are used to calculate Index scores for domains of immediate memory, visuospatial/constructional, attention, language, delayed memory, and global neuropsychological functioning. The index scores utilize age-corrected normative comparisons from the test manual to generate standard scores (M = 100, SD = 15). Higher scores indicating better cognition.
- The 30-item GDS (Yesavage et al., 1982) was used to assess self-reported depression for the first sample, and the 15-item GDS (Sheikh & Yesavage, 1986) was used to assess self-reported depression for the second sample. Higher scores indicated more self-reported depression for both measures, with the second sample using a cut-off of 5/15 (or higher) as exclusion for the parent study.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

After approximately one week, the HVLT-R, BVMT-R, SDMT, TMT-A, and TMT-B were repeated. The same form of each test was used to maximize practice effects for these repeated cognitive tasks. These tasks represented the primary measures in the current study. The RBANS and WRAT were only administered at baseline, and participants were classified as being cognitively intact based on their performance on baseline scores from all tests. The GDS was also administered at baseline. With the exception of SDMT possessing one missing value, no other variables of interest possessed missing data.

Analyses

Pairwise Baseline Versus One-Week Analyses—Pair-wise *t* tests were conducted to compare Observed Baseline and Observed One-Week scores for each of the repeated measures in the cognitive battery (HVLT-R, BVMT-R, SDMT, TMT-A, TMT-B) to approximate a traditional evaluation of change over time (comparison of Time 1 and Time 2 scores) without controlling for practice effects or participant variables.

SRB Group Analyses—Previously published SRB prediction equations for each of the measures in the cognitive battery were applied to the current combined validation sample's Baseline and One-Week scores (see Table 2 for Duff's 2014 SRB equations). As has been described previously (Duff, 2014), the SRB prediction algorithms were calculated from a development sample using stepwise multiple-regression analyses to maximize the prediction of performance for each repeated measure in the cognitive battery. Demographic variables (e.g., age, education, sex), and baseline test score were used to predict the respective test score at one-week follow-up.

Following the application of these SRB prediction equations to the current intact combined validation sample, a z score was calculated for each participant. This z score reflects a normalized deviation of change for an individual participant. Specifically, the Observed One-Week score (T_2) was compared to the Predicted One-Week score (T_2'), normalized by the SE_{est} (i.e., $z = (T_2 - T_2')/ SE_{est}$). While some discussion in the literature exists regarding the proper standard error estimate for use in reliable change methods (Hinton-Bayre, 2010), we have previously shown the equivalence of the two most-common estimates and provided support for use of the SE_{est} (Hammers & Duff, 2019). Z scores for each repeated measure were then compared to expectation ($z = 0$) based on the normal distribution of z scores using a one-sample t test.

Individual Distribution Analyses—The resultant z scores were additionally trichotomized into "smaller-than-expected practice effects" (z score < −1.645), "expected practice effects" (z score falling between +/− 1.645), or "greater-than-expected practice effects" (z score > 1.645) for all measures in the repeated battery, with the exception of TMT-A and TMT-B that used reversed scoring. As indicated previously, if the z scores were normally distributed, then it would be expected that 5% of participants show "smaller-than-expected practice effects," 90% would show "expected practice effects", and 5% would reflect "greater-than-expected practice effects." Using this trichotomization, individual one-sample chi-square analyses were conducted for each measure in the repeated cognitive

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

battery to determine if the observed distribution of participants' performance deviated significantly from the expected distribution based on the normal distribution of *z* scores.

Calculation of SRBs Developed from the Current Combined Development Sample—Finally, multiple regression analyses were used to derive updated prediction equations from the pooling of the current combined validation sample of 107 cognitively intact participants with the 93 cognitively intact participants from Duff's (2014) development sample, resulting in a total combined development sample of 200 cognitively intact participants. The updated prediction equations were based on McSweeny et al.'s (1993) method, though the current study used hierarchical multiple regression instead of step-wise multiple regression analyses for better statistical rigor (Millis, 2003). A separate prediction equation was generated for each of the cognitive measures in the repeated test battery. Similar to the methodology used by Duff (2014) and others (Beatty, Mold, & Gontkovsky, 2003; Patton et al., 2003), demographic variables (age, education, sex, premorbid intellect), retest interval, and the Time 1 (T_1) score were regressed on the respective Time 2 (T_2) score, though in the current sample the predictors were regressed in separate blocks (as compared to in a step-wise fashion). Specifically, Block 1 contained T_1 score, Block 2 added age, Block 3 added education, Block 4 added sex, Block 5 added premorbid intellect, and Block 6 added retest interval. For example, demographic variables, retest interval, and T_1 HVLT-R Total Recall was regressed on T_2 HVLT-R Total Recall performance. Age and education were represented as years old at T_1 and number of years of formal education, respectively. Premorbid intellect was measured by WRAT Reading subtest performance. Sex was coded as *male* = 0, and fe*male* = 1. The retest interval was represented as days from T_1 to T_2. All scores used in the SRB equations were raw scores, with the exception of premorbid intellect (standard scores).

Measures of effect size were expressed throughout as Cohen's *d* values for continuous data, and *Phi* coefficients for categorical data. Given the number of comparisons in the current study, a two-tailed alpha level was set at .01 for all statistical analyses.

RESULTS

Pairwise Baseline Versus One-Week Analyses

Change over time was first assessed using a traditional method of comparing Observed Baseline and Observed One-Week scores for each of the repeated measures in the cognitive battery (HVLT-R, BVMT-R, SDMT, TMT-A, TMT-B; see Table 3 for means) in these intact combined validation samples of community-dwelling older adults. Significant differences were observed for HVLT-R Total Recall, $t(106) = -8.95$, $p = .001$, $d = -1.74$, HVLT-R Delayed Recall, $t(106) = -6.68$, $p = .001$, $d = -1.30$, BVMT-R Total Recall, $t(106) = -19.37$, $p = .001$, $d = -3.76$, BVMT-R Delayed Recall, $t(106) = -8.67$, $p = .001$, $d = -1.68$, SDMT, $t(105) = -6.07$, $p = .001$, $d = -1.18$, and TMT-B, $t(106) = 3.45$, $p = .001$, $d = 0.67$. Specifically, scores were better at Observed One-Week than at Observed Baseline for all six measures. No difference was observed for the TMT-A, $t(106) = 0.71$, $p = .48$, $d = 0.13$.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

SRB Group Analyses

Duff's (2014) SRB prediction equations for each of the repeated measures in the cognitive battery were applied to the current combined validation sample. When using one-sample *t* tests to compare *z* scores for each repeated measure to expectation ($z = 0$) based on the normal distribution of z scores (Table 3), the *z* score for BVMT-R Total Recall was significantly lower than zero, $t(106) = -2.83$, $p = .006$, $d = -0.55$. As a reminder, a negative z score means that Observed One-Week score is lower than the Predicted One-Week score, such that the current combined validation sample fell below expectations on this task based on Duff's development sample. Conversely, no significant differences were observed on any of the other measures administered twice over one week, HVLT-R Total Recall, $t(106) = -0.88$, $p = .38$, $d = -0.17$, HVLT-R Delayed Recall, $t(106) = -1.48$, $p = .14$, $d = -0.29$, BVMT-R Delayed Recall, $t(106) = -0.12$, $p = .91$, $d = -0.02$, SDMT, $t(105) = -0.17$, $p = .87$, $d = -0.03$, TMT-A, $t(106) = 1.56$, $p = .12$, $d = 0.30$, or TMT-B, $t(106) = -0.75$, $p = .46$, $d = -0.15$.

Individual Distribution Analyses

Next, we examined the distribution of individual intact older adults that displayed "smaller-than-expected practice effects" (*z* score < −1.645 for HVLT-R, BVMT-R, and SDMT; *z* score > 1.645 for TMT-A and TMT-B), "expected practice effects" (*z* score falling between +/− 1.645), or "greater-than-expected practice effects" (*z* score > 1.645 for HVLT-R, BVMT-R, and SDMT; *z* score < −1.645 for TMT-A and TMT-B) between Baseline and One-Week administrations of the repeated cognitive battery. The majority of participants exhibited the expected level of improvement or practice effect (94.6% of participants; see Table 4). Similarly, no significant difference in performance distribution was seen relative to expectation for any of the cognitive measures in this repeated battery, HVLT-R Total Recall, χ^2 (2) = 2.26, $p = .32$, *Phi* = .14, BVMT-R Total Recall, χ^2 (2) = 5.70, $p = .06$, *Phi* = .23, SDMT, χ^2 (2) = 0.36, $p = .83$, *Phi* = .06, TMT-A, χ^2 (2) = 2.67, $p = .26$, *Phi* = .16, or TMT-B, χ^2 (2) = 3.47, $p = .18$, *Phi* = .18. Although non-significant trends were observed for HVLT-R Delayed Recall, χ^2 (2) = 7.62, $p = .02$, *Phi* = .27, and BVMT-R Delayed Recall, χ^2 (2) = 6.25, $p = .04$, *Phi* = .24, these suggested that more people displayed performances over one week that *were consistent* with Duff's (2014) prediction equations than was anticipated based on the normal distribution of *z* scores.

Calculation of SRBs Developed from the Current Combined Development Sample

As described above, prediction equations for T_2 scores were calculated for each measure in the repeated test battery, based on pooling the 107 cognitively intact participants from the current combined validation sample and the 93 cognitively intact participants from Duff's (2014) development sample (for a total $n = 200$ for the combined development sample). The results of the prediction equations are presented in Table 5. For each score, the final model's R^2, SE_{est}, constant, and unstandardized beta weights for relevant variables are listed. The final model predicting HVLT-R Total Recall at T_2, which included HVLT-R Total Recall at T_1, accounted for a significant amount of variance, $F(1, 198) = 151.46$, $p < .001$. HVLT-R Delayed Recall T_2 was best predicted by HVLT-R Delayed Recall T_1, $F(1, 198) = 162.54$, $p < .001$. BVMT-R Total Recall T_2 was best predicted by BVMT-R Total Recall T_1, age,

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

education, sex, and WRAT Reading, $F(5, 194) = 49.33$, $p < .001$. BVMT-R Delayed Recall T_1, age, education, sex, and WRAT Reading best predicted BVMT-R Delayed Recall T_2, $F(5, 194) = 46.10$, $p < .001$. SDMT T_2 was best predicted by SDMT T_1, age, and education, $F(3, 195) = 168.19$, $p < .001$. TMT-A T_1 best predicted TMT-A T_2, $F(1, 198) = 64.13$, $p < .001$. Finally, the follow-up TMT-B score was best predicted by TMT-B T_1, age, education, sex, WRAT Reading, and retest interval, F(6, 193) = 45.50, $p < .001$.

DISCUSSION

The current study sought to examine the validity of previously published SRB predicted difference equations (Duff, 2014) for commonly administered cognitive measures (HVLT-R, BVMT-R, SDMT, TMT-A, and TMT-B) in an independent sample of cognitively intact community-dwelling older adults assessed twice over a one-week period. While a few studies (Duff et al., 2018; Duff et al., 2017; Hammers, Suhrie, et al., 2020) have previously attempted to externally validate these SRB prediction equations in predominantly impaired samples, to our knowledge, this is the first study to apply these equations to a large and clean sample of cognitively intact participants. Use of cognitively-intact validation samples permits the opportunity to examine whether Duff's use of cognitively compromised participants in its development sample (e.g., 44% of Duff's sample was MCI) leads to under-prediction of the SRB prediction equations in intact individuals. We also attempted to update Duff's (2014) prediction equations by increasing the size of the combined development sample ($n = 200$), including only cognitively intact participants, and adding select participant and test characteristic variables to the algorithms – premorbid intellect and retest interval – based on their known impact on practice effects.

For our current combined validation sample of cognitively intact participants, when comparing observed test scores at baseline and one-week, large and statistically significant improvements in performance were observed across most measures administered (HVLT-R Total Recall, HVLT-R Delayed Recall, BVMT-R Total Recall, BVMT-R Delayed Recall, TMT-B, and SDMT; Cohen's *d*s = | 0.67 – 3.76 |). For example, as seen in Table 3, BVMT-R Total Recall improved, on average, from 21.5 points at baseline to 29.5 points at one-week follow-up, accounting for a 36.8% improvement. Given the short test-retest interval and the lack of expected acute change in cognition for this sample over one week, these improved performances are being interpreted as a practice effect. These results are consistent with the multitude of research (Darby, Maruff, Collie, & McStephen, 2002; Duff et al., 2018; Duff et al., 2017; Fernandez-Ballesteros, Zamarron, & Tarraga, 2005; Hammers, Suhrie, et al., 2020; Suchy, Kraybill, & Franchow, 2011) showing that older adults exhibit practice effects on repeat testing. Specifically for BVMT-R, this visual memory task appears to be especially susceptible to benefit from repeated exposure, which coincides with a meta-analysis by Calamia et al. (2012) showing that the largest positive practice effects observed for a cognitive domain was for visual memory.

Conversely, when applying Duff's (2014) SRB prediction equations to baseline performance on these measures, our combined validation sample's level of observed practice effect was consistently *within expectation* of predictions across nearly all measures administered. This was observed in two different methods. First, when considering group-level analysis, only

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

one significant difference was observed between predicted one-week scores and observed one-week scores for our combined validation sample of intact participants (BVMT-R Total Recall; see Table 3). Second, when examining the distributions of individual participants in our combined validation sample that displayed smaller-than-expected, expected, or greater-than-expected practice effects, no significant differences in performance distribution were seen relative to expectation for any of the cognitive measures in this repeated battery, and some trends were observed where more participants displayed performance over one week that *was consistent* with Duff's (2014) prediction equations than was anticipated based on the normal distribution of *z* scores (i.e., HVLT-R Delayed Recall and BVMT-R Delayed Recall). As a reminder, the properties of the normal curve for a 90% CI at an $\alpha = 0.05$ indicate that 5% of participants should display smaller-than-expected practice effects, 90% should show expected practice effects, and 5% should reflect greater-than-expected practice effects. As seen in Table 4, however, few participants possessed worse-than or better-than expected practice effects based on normal distributions, with on average 95% of the individual participants displaying the expected level of benefit from practice.

These collective results are counter to our expectation based on Duff's (2014) use of a development sample that included both cognitive intact participants and those with MCI. Specifically, it was anticipated that Duff's inclusion of cognitively impaired participants in the development sample would have reduced the purity of that sample, led to differential rates of change relative to intact samples, and subsequently impacted the ability of the SRB equations to predict performance at Time 2 when applied to non-impaired samples. This concern is related to the importance in clinical neuropsychology of normative (and normative change) samples being cognitively intact (Goodwill et al., 2019; Harrington et al., 2017), as compared to being from a mixed sample like in Duff (2014). Instead, Duff's prediction equations appear to predict performance at One-Week well in this validation sample, and the only measure where predicted performance differed from observed performance (BVMT-R Total Recall) suggested that the prediction equations *over-predicted* performance. As expected given the presence of individuals with MCI, Duff's (2014) sample displayed consistently lower baseline scores than our current intact sample on the respective measures (HVLT-R Total Recall = 23.2 for Duff vs. 27.5 currently; HVLT-R Delayed Recall= 6.7 for Duff vs. 9.8 currently; BVMT-R Total Recall = 14.6 for Duff vs. 21.5 currently; BVMT-R Delayed Recall = 5.6 for Duff vs. 8.9 currently; TMT-A = 44.1 for Duff vs. 36.5 currently; TMT-B = 117.1 for Duff vs. 85.6 currently; and SDMT = 39.5 for Duff vs. 44.5 currently). Normatively, Duff's samples tended to fall, on average, in the low average to average range, and the current combined validation sample participants were consistently in the average range relative to their same-aged peers. The two samples (Duff's and the current combined validation sample) were relatively equal with respect to age (78.6 years for Duff vs. 73.3 years currently), education (15.4 years for Duff vs. 16.1 years currently), premorbid intellect (107.8 *SS* for Duff vs. 109.6 *SS* currently), retest interval (Duff reported at "one-week" vs. 7.3 days currently), and both sex (81.1% female for Duff vs. 72.0% currently) and ethnicity (100% Caucasian for Duff vs. 99.0% Caucasian currently). Conversely, when comparing the benefit from repeat test exposure – practice effect – between the two samples, Duff's sample tended to benefit from practice to generally comparable levels (HVLT-R Total Recall = 4.4 points improvement for Duff vs. 3.1

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

currently; HVLT-R Delayed Recall= 2.2 points improvement for Duff vs. 0.9 currently; BVMT-R Total Recall = 8.2 points improvement for Duff vs. 7.9 currently; BVMT-R Delayed Recall = 2.3 points improvement for Duff vs. 1.4 currently; TMT-A = 4.2 seconds improvement for Duff vs. 1.3 currently; TMT-B = 12.4 seconds improvement for Duff vs. 9.8 currently; and SDMT = 2.3 points improvement for Duff vs. 2.8 currently) despite 44% of their sample possessing a diagnosis of MCI. While the current combined validation sample may have potentially been limited by ceiling effects at Time 2 given their stronger Time 1 performances, this general comparability in practice effects between the samples likely explains the ability of Duff's SRB prediction equations to predict Time 2 performance in our sample with accuracy, and potentially sheds light on the relative contribution of the various components of complex-SRB regression equations. It is also worth noting that Duff's (2014) mixed sample likely allowed for a wider range of test performances than a cognitively intact sample alone, which enhances the generalizability of regression models. Taken together, the current results appear to add further external support to the validity of Duff's (2014) SRB equations for the cognitive measures administered to predict cognitive performance in research or clinical situations when repeated administration of testing is necessary over a short period of time.

The current study additionally updated Duff's (2014) prediction equation models by pooling the 107 participants from the current combined validation sample with the 93 cognitively intact participant's from Duff's original development sample to create a combined development sample (n = 200), and examined the impact of adding select participant and test characteristic variables as predictors to the models. As can be observed in Table 5, when incorporating baseline performance, demographic variables (age, education, sex, premorbid intellect), and retest interval, we were able to significantly predict T_2 performance for each of the measures in our repeated test battery. WRAT premorbid intellect was a significant predictor in 3 of the 7 scores (BVMT-R Total Recall, BVMT-R Delayed Recall, and TMT-B) examined, and retest interval was a significant predictor for one measure examined (TMT-B). These results are consistent with past research suggesting strong associations between greater premorbid intellect and stronger practice effects (Patton et al., 2005; Rapport et al., 1997; Rapport et al., 1997), along was previous findings that premorbid intellect has been shown to play a moderating effect in the relationship between β-amyloid accumulation and cognition (Duff et al., 2013). As premorbid intellect has been used as a proxy measure for cognitive reserve (Jefferson et al., 2011), it has been suggested as a protective factor in the development of dementia (Stern, 2006), an influence on the onset of AD symptoms (Roe, Xiong, Grant, Miller, & Morris, 2008), and a moderator in the development of AD pathology (Rentz et al., 2010; Rodrigue et al., 2012; Roe, Mintun, et al., 2008). Relatedly, our results are consistent with previous findings by Calamia and colleagues (2012) showing that the interval between repeated test administrations may have significant impacts on the effect of practice during serial administration.

When comparing the R^2 and SE_{est} values between the two sets of prediction equations (see Table 2 for Duff's 2014 and Table 5 for the current equations), some trends tended to emerge. For all measures but TMT-A, the R^2 value, or the amount of variance accounted for in the model, was relatively comparable between the two sets of equations. In fact, the difference in the amount of variance between Duff's and the current models were only 2 –

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

15% (0.15 R^2 units) of each other. Some tasks were predicted slightly better by Duff's equations (HVLT-R Total Recall, HVLT-R Delayed Recall, BVMT-R Total Recall, BVMT-R Delayed Recall, and SDMT) and others by the current equations (TMT-B). Based on this metric, neither set of prediction equations was notably better than the other.

Of greater importance, however, the current SRB prediction equations resulted in a smaller SE_{est} value for every task administered in the repeated battery except TMT-A (commonly ranging from a 28 – 38% reduction in the SE_{est}). This reduction is important because, as a reminder, the calculation of z scores for reliable change methods relies on the SE_{est} to determine if a significant deviation exists between an individual's predicted and observed Time 2 performance. Specifically, when considering the equation, $z = (T_2 - T_2')/ SE_{est}$, prediction equations generating smaller SE_{est} values will mean a smaller difference between an individual's predicted and observed Time 2 performance will be necessary for a given z score, or significance at a set cut-off. A case example may be of benefit to better explain the differences between the SE_{est} derived from change formulae calculated from Duff's (2014) development sample and the current pooled sample. Imagine that a 75 year-old female with 12 years of education was tested twice across one week on the HVLT-R Total and Delayed Recall, and her Observed Baseline and Follow-up scores were as follows: Observed Baseline HVLT-R Total Recall = 30 vs. Observed Follow-up HVLT-R Total Recall = 27, and Observed Baseline HVLT-R Delayed Recall = 10 vs. Observed Follow-up HVLT-R Delayed Recall = 8. She declines slightly on both measures by 2–3 points over the two test administrations. As can be seen in Table 6, when applying both Duff's (2014) and the currently updated SRB equations, it is observed that her Predicted scores should be similar regardless of the development sample and set of equations. However, the currently updated SRBs' smaller SE_{est} (HVLT-R Total Recall $SE_{est} = 3.71$ for Duff vs. 3.34 currently, HVLT-R Delayed Recall $SE_{est} = 1.88$ for Duff vs. 1.17 currently) means that the relatively same deviation between Observed and Predicted scores between the two sets of SRBs leads to different conclusions about her performance. More specifically, despite these variations from Time 1 to Time 2, neither of these performances suggest a statistically significant deviation from expectation based on Duff's (2014) SRB prediction equations, as the z scores are all <1.645. Conversely, both performances reflect a statistically significant decline relative to expectation when using the current SRB prediction equations (HVLT-R Total and Delayed Recall $zs = -1.69$ and -2.39, respectively). This difference in SE_{est} is not necessarily surprising given that Duff's sample incorporated both cognitively intact participants and those with MCI, resulting in a wider potential range of performance compared to our cleanly cognitively intact sample, though our increased sample size in the updated SRBs from pooling across three samples could have theoretically inflated our degree of deviation as well. Overall, these results coincide with the primary conclusion in the current study that Duff's SRB equations tended to accurately predict performance across the current cognitively intact sample, but also highlight the importance of premorbid intellect and retest interval in predicting follow-up performance during serial assessment, and using cognitively intact normative samples.

An interesting observation from reviewing Table 5 was that improved prediction of T_2 performance by demographic variables was not uniform across the newly-created SRB equations. It appears that some cognitive measures are more sensitive to demographic input

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

than others, as can been seen by the variables of age, education, sex, and WRAT premorbid intellect significantly predicting T_2 performance on the BVMT-R measures but not the HVLT-R measures. This can also be observed, to a certain extent, in the SRB equations from Duff's (2014) original development sample (see Table 2), such that demographic variables predicted T_2 scores for HVLT-R to a lesser extent than BVMT-R. It has been proposed that BVMT-R has greater sensitivity to the AD continuum than does HVLT-R, given that BVMT-R has previously shown stronger relationships with the AD biomarker hippocampal volume than HVLT-R (Duff et al., 2018; Hammers, Kucera, et al., 2020). As age, and to a certain extent demographic variables of education, sex, and premorbid intellect, have been shown to be predictive of both cognition and the development of AD (Salthouse, 2009), it would therefore not be surprising that these variables are more highly predictive of BVMT-R future performance than HVLT-R future performance in the current study.

The current study is not without limitations. These findings are specific to the cognitive measures administered in this battery over this particular time frame (one week), and generalization cannot be made to other measures of cognition, different retest intervals, or use of alternative forms (as shown by Calamia and colleagues; Calamia et al., 2012). Future studies should build upon previous work of Duff and colleagues (Duff, Beglinger, Moser, Paulsen, et al., 2010) to show that SRBs may transcend specific tests within a domain or retest intervals. Also, while the range of retest intervals was restricted to 4 – 14 days during the current study, the focus of the manuscript was on short-term practice effects therefore this range (10 days) was relatively broad given the overall mean retest interval of 7.3 days. With older research indicating that practice effects for cognitive testing were greatest at 1 week and declined for intervals of 1 month, 2 months, or 4 months (Catron & Thompson, 1979) – and the goal of these SRBs was to predict Time 2 performance optimally with data generally available to researchers and clinicians – we felt that this 10-day range of retest interval was appropriate for inclusion in the prediction equations. In addition, these results may not generalize to more heterogeneous participants in regards to premorbid functioning, sex, education, and ethnicity. While future research should consider such predictions in samples that are not primarily well-educated Caucasian females, the current study's proportions of highly educated females in each sample (collected in different states and over one-decade apart) appear to reflect long-standing trends in research participation. Specifically, it has been observed that women tend to volunteer more than men across all age ranges (United States Bureau of the Census, Statistics, National, & Service, 2015), reaching a difference of upwards of 30% (U.S. Bureau of Labor Statistics, 2016), and that individuals with higher incomes and greater levels of education consistently volunteer at greater levels (United States Bureau of the Census et al., 2015). Future SRB studies should therefore be more intentional at recruiting samples more reflective of the general population. Further, as the specific focus of this manuscript was to evaluate Duff's (2014) SRB performances in an intact sample, the current results do not provide information on cognitive prediction in other disease states.

Finally, performance by our intact participants in the current combined validation sample on the TMT-A was inconsistent with the remainder of the measures in the repeated test battery across nearly all analyses conducted. For example, TMT-A performance failed to display significant improvement (or practice effect) from Time 1 to Time 2, and prediction of

Author Manuscript

follow-up score was not improved like most other measures with the inclusion of premorbid intellect or retest interval (see R^2 and SE_{est} values). Our previous work (Hammers, Suhrie, et al., 2020) showed that older adults with MCI displayed improved performance over one week on this measure of visual scanning and speeded processing, rendering the current results surprising in the cognitively intact sample (that should be more prone to benefitting from practice). When reviewing mean and standard deviations for the measure across the two administrations in Table 3, at least a subset of cognitively intact participants completing this task displayed a high degree of variability at Time 2 (*SD* of 13.7 seconds at Time 1 vs. 20.2 seconds at Time 2). This is in contrast to the same sample's performance on the more challenging TMT-B, where significant practice effect was observed across the sample at Time 2, and the *SD* declined from 39.5 seconds for Time 1 to 31.6 seconds for Time 2. Further examination of this TMT-A variability is warranted in future studies.

Despite these limitations, the results support the validity of Duff's (2014) SRB prediction equations, the first study to do so in a cognitively intact sample. They also support the potential for these one-week SRB prediction equations to possess value diagnostically and prognostically, and inform treatment recommendations given research suggesting that practice effects can predict response to intervention (Duff, Beglinger, Moser, Schultz, & Paulsen, 2010), outcomes (Duff et al., 2011; Hassenstab et al., 2015; Machulda et al., 2013), and Alzheimer's-related pathology (Duff et al., 2018; Duff, Foster, & Hoffman, 2014; Galvin et al., 2005; Mormino et al., 2014). Using short-term practice effects as a screening tool in integrated primary care settings has the potential to inform clinical decision making about individuals along the Alzheimer's continuum, and accelerate the time to possible intervention.

Acknowledgments

Funding: The project described was supported by research grants from the National Institutes on Aging: 5R01AG045163. 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

ADNI2. (2020). Alzheimer's Disease Neuroimaging Initiative: ADNI2 Procedures Manual. Retrieved from https://adni.loni.usc.edu/wp-content/uploads/2008/07/adni2-procedures-manual.pdf

Attix DK, Story TJ, Chelune GJ, Ball JD, Stutts ML, Hart RP, & Barth JT (2009). The prediction of change: normative neuropsychological trajectories. Clin Neuropsychol, 23(1), 21–38. doi:10.1080/13854040801945078 [PubMed: 18720272]

Beatty WW, Mold JW, & Gontkovsky ST (2003). RBANS performance: influences of sex and education. J Clin Exp Neuropsychol, 25(8), 1065–1069. doi:10.1076/jcen.25.8.1065.16732 [PubMed: 14566580]

Benedict R. (1997). Brief visuospatial memory test - Revised: professional manual. Lutz, FL: Psychological Assessment Resources, Inc.

Brandt J, & Benedict R. (2001). Hopkins Verbal Learning Test - Revised. Odessa, FL: PAR.

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/pubmed/22540222 [PubMed: 22540222]

Catron DW, & Thompson CC (1979). Test-retest gains in WAIS scores after four retest intervals. J Clin Psychol, 35(2), 352–357. doi:10.1002/1097-4679(197904)35:2<352::aid-jclp2270350226>3.0.co;2-2 [PubMed: 457898]

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Crockford C, Newton J, Lonergan K, Madden C, Mays I, O'Sullivan M, . . . Abrahams S. (2018). Measuring reliable change in cognition using the Edinburgh Cognitive and Behavioural ALS Screen (ECAS). Amyotroph Lateral Scler Frontotemporal Degener, 19(1–2), 65–73. doi:10.1080/21678421.2017.1407794 [PubMed: 29214872]

Darby D, Maruff P, Collie A, & McStephen M. (2002). Mild cognitive impairment can be detected by multiple assessments in a single day. Neurology, 59(7), 1042–1046. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12370459 [PubMed: 12370459]

Duff K. (2014). One-week practice effects in older adults: tools for assessing cognitive change. Clin Neuropsychol, 28(5), 714–725. doi:10.1080/13854046.2014.920923 [PubMed: 24882553]

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, Paulsen JS, Schultz SK, & Arndt S. (2010). Predicting cognitive change in older adults: the relative contribution of practice effects. Arch Clin Neuropsychol, 25(2), 81–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20064816 [PubMed: 20064816]

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, Foster NL, Dennett K, Hammers DB, Zollinger LV, Christian PE, . . . Hoffman JM (2013). Amyloid deposition and cognition in older adults: the effects of premorbid intellect. Arch Clin Neuropsychol, 28(7), 665–671. doi:10.1093/arclin/act047 [PubMed: 23817438]

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. doi:10.1097/JGP.0b013e318209dd3a [PubMed: 22024617]

Duff K, Schoenberg MR, Patton D, Mold J, Scott JG, & Adams RL (2004). Predicting change with the RBANS in a community dwelling elderly sample. J Int Neuropsychol Soc, 10(6), 828–834. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15637773 [PubMed: 15637773]

Duff K, Schoenberg MR, Patton D, Paulsen JS, Bayless JD, Mold J, . . . Adams RL (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. Arch Clin Neuropsychol, 20(3), 281–290. doi:10.1016/j.acn.2004.07.007 [PubMed: 15797165]

Fernandez-Ballesteros R, Zamarron MD, & Tarraga L. (2005). Learning potential: a new method for assessing cognitive impairment. Int Psychogeriatr, 17(1), 119–128. doi:10.1017/s1041610205000992 [PubMed: 15945596]

Folstein MF, Folstein SE, & McHugh PR (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res, 12(3), 189–198. doi:10.1016/0022-3956(75)90026-6 [PubMed: 1202204]

Galvin JE, Powlishta KK, Wilkins K, McKeel DW Jr., Xiong C, Grant E, . . . Morris JC (2005). Predictors of preclinical Alzheimer disease and dementia: a clinicopathologic study. Arch Neurol, 62(5), 758–765. doi:10.1001/archneur.62.5.758 [PubMed: 15883263]

Gavett BE, Ashendorf L, & Gurnani AS (2015). Reliable Change on Neuropsychological Tests in the Uniform Data Set. J Int Neuropsychol Soc, 21(7), 558–567. doi:10.1017/S1355617715000582 [PubMed: 26234918]

Goodwill AM, Campbell S, Henderson VW, Gorelik A, Dennerstein L, McClung M, & Szoeke C. (2019). Robust norms for neuropsychological tests of verbal episodic memory in Australian women. Neuropsychology, 33(4), 581–595. doi:10.1037/neu0000522 [PubMed: 30829514]

Hammers D, Duff K, & Chelune G. (2015). Assessing change of cognitive trajectories over time in later life In Pachana NA & Laidlaw K (Eds.), Oxford Handbook of Clinical Geropsychology. Oxford: Oxford University Press.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hammers DB, & Duff K. (2019). Application of Different Standard Error Estimates in Reliable Change Methods. Arch Clin Neuropsychol. doi:10.1093/arclin/acz054

Hammers DB, Kucera A, Spencer RJ, Abildskov TJ, Archibald ZG, Hoffman JM, & Wilde EA (2020). Examining the Relationship between a Verbal Incidental Learning Measure from the WAIS-IV and Neuroimaging Biomarkers for Alzheimer's Pathology. Dev Neuropsychol, 45(3), 95–109. doi:10.1080/87565641.2020.1762602 [PubMed: 32374196]

Hammers DB, Suhrie KR, Dixon A, Porter S, & Duff K. (2020). Reliable change in cognition over 1 week in community-dwelling older adults: a validation and extension study. Arch Clin Neuropsychol. doi:10.1093/arclin/acz076

Harrington KD, Lim YY, Ames D, Hassenstab J, Rainey-Smith S, Robertson J, . . . Group, A. R. (2017). Using Robust Normative Data to Investigate the Neuropsychology of Cognitive Aging. Arch Clin Neuropsychol, 32(2), 142–154. doi:10.1093/arclin/acw106 [PubMed: 27932344]

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]

Hinton-Bayre AD (2010). Deriving reliable change statistics from test-retest normative data: comparison of models and mathematical expressions. Arch Clin Neuropsychol, 25(3), 244–256. doi:10.1093/arclin/acq008 [PubMed: 20197293]

Jefferson AL, Gibbons LE, Rentz DM, Carvalho JO, Manly J, Bennett DA, & Jones RN (2011). A life course model of cognitive activities, socioeconomic status, education, reading ability, and cognition. J Am Geriatr Soc, 59(8), 1403–1411. doi:10.1111/j.1532-5415.2011.03499.x [PubMed: 21797830]

Lezak M, Howieson D, Bigler E, & Tranel D. (2012). Neuropsychological Assessment (5th. ed.). New York: Oxford University Press.

Machulda MM, Pankratz VS, Christianson TJ, Ivnik RJ, Mielke MM, Roberts RO, . . . Petersen RC (2013). Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging. Clin Neuropsychol, 27(8), 1247–1264. doi:10.1080/13854046.2013.836567 [PubMed: 24041121]

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. The Clinical Neuropsychologist 7, 300–312.

Millis S. (2003). Statistical practices: the seven deadly sins. Child Neuropsychol, 9(3), 221–233. doi:10.1076/chin.9.3.221.16455 [PubMed: 13680411]

Mormino EC, Betensky RA, Hedden T, Schultz AP, Amariglio RE, Rentz DM, . . . Sperling RA (2014). Synergistic effect of beta-amyloid and neurodegeneration on cognitive decline in clinically normal individuals. JAMA Neurol, 71(11), 1379–1385. doi:10.1001/jamaneurol.2014.2031 [PubMed: 25222039]

Morris JC (1993). The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology, 43(11), 2412–2414. doi:10.1212/wnl.43.11.2412-a

Patton DE, Duff K, Schoenberg MR, Mold J, Scott JG, & Adams RL (2003). Performance of cognitively normal African Americans on the RBANS in community dwelling older adults. Clin Neuropsychol, 17(4), 515–530. doi:10.1076/clin.17.4.515.27948 [PubMed: 15168916]

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]

Randolph C. (2012). Repeatable Battery for the Assessment of Neuropsychological Status. Bloomington, MN: The Psychological Corporation.

Rapport LJ, Axelrod BN, Theisen ME, Brines DB, Kalechstein AD, & Ricker JH (1997). Relationship of IQ to verbal learning and memory: test and retest. J Clin Exp Neuropsychol, 19(5), 655–666. doi:10.1080/01688639708403751 [PubMed: 9408796]

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

Reitan R. (1992). Trail Making Test: Manual for administration and scoring. Tucson, AZ: Reitan Neuropsychology Laboratory.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Rentz DM, Locascio JJ, Becker JA, Moran EK, Eng E, Buckner RL, . . . Johnson KA (2010). Cognition, reserve, and amyloid deposition in normal aging. Ann Neurol, 67(3), 353–364. doi:10.1002/ana.21904 [PubMed: 20373347]

Rinehardt E, Duff K, Schoenberg M, Mattingly M, Bharucha K, & Scott J. (2010). Cognitive change on the repeatable battery of neuropsychological status (RBANS) in Parkinson's disease with and without bilateral subthalamic nucleus deep brain stimulation surgery. Clin Neuropsychol, 24(8), 1339–1354. doi:10.1080/13854046.2010.521770 [PubMed: 20967688]

Rodrigue KM, Kennedy KM, Devous MD Sr., Rieck JR, Hebrank AC, Diaz-Arrastia R, . . . Park, D. C. (2012). beta-Amyloid burden in healthy aging: regional distribution and cognitive consequences. Neurology, 78(6), 387–395. doi:10.1212/WNL.0b013e318245d295 [PubMed: 22302550]

Roe CM, Mintun MA, D'Angelo G, Xiong C, Grant EA, & Morris JC (2008). Alzheimer disease and cognitive reserve: variation of education effect with carbon 11-labeled Pittsburgh Compound B uptake. Arch Neurol, 65(11), 1467–1471. doi:10.1001/archneur.65.11.1467 [PubMed: 19001165]

Roe CM, Xiong C, Grant E, Miller JP, & Morris JC (2008). Education and reported onset of symptoms among individuals with Alzheimer disease. Arch Neurol, 65(1), 108–111. doi:10.1001/archneurol.2007.11 [PubMed: 18195147]

Salthouse TA (2009). Decomposing age correlations on neuropsychological and cognitive variables. J Int Neuropsychol Soc, 15(5), 650–661. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19570312 [PubMed: 19570312]

Sanchez-Benavides G, Pena-Casanova J, Casals-Coll M, Gramunt N, Manero RM, Puig-Pijoan A, . . . Team, N. S. (2016). One-Year Reference Norms of Cognitive Change in Spanish Old Adults: Data from the NEURONORMA Sample. Arch Clin Neuropsychol, 31(4), 378–388. doi:10.1093/arclin/acw018 [PubMed: 27193368]

Sheikh JI, & Yesavage J. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. Clinical Gerontologist, 5, 165–172.

Smith A. (1973). Symbol Digit Modalities Test. Los Angeles, CA: Western Psychological Services.

Stein J, Luppa M, Brahler E, Konig HH, & Riedel-Heller SG (2010). The assessment of changes in cognitive functioning: reliable change indices for neuropsychological instruments in the elderly - a systematic review. Dement Geriatr Cogn Disord, 29(3), 275–286. doi:10.1159/000289779 [PubMed: 20375509]

Stern Y. (2006). Cognitive reserve and Alzheimer disease. Alzheimer Dis Assoc Disord, 20(2), 112–117. doi:10.1097/01.wad.0000213815.20177.19 [PubMed: 16772747]

Suchy Y, Kraybill ML, & Franchow E. (2011). Practice effect and beyond: reaction to novelty as an independent predictor of cognitive decline among older adults. J Int Neuropsychol Soc, 17(1), 101–111. doi:10.1017/S135561771000130X [PubMed: 21073771]

U.S. Bureau of Labor Statistics. (2016). Volunteering in the United States, 2015 [Press release]. Retrieved from https://www.bls.gov/news.release/volun.nr0.htm

United States Bureau of the Census, Statistics, U. S. D. o. L. B. o. L., National, C. f., & Service, C. (2015). Current Population Survey, 9 2014: Volunteer Supplement.

Wechsler D. (1987). Manual for the Weschler Memory Scale - Revised. San Antonio, Tx: The Psychological Corporation.

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]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Demographic characteristics of the current validation samples, and the combined validation and development samples.

	First Validation Sample	**Second Validation Sample**	**Combined Validation Sample**	**Combined Development Sample**
Variable	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
n	55	52	107	200
Age (years)	74.1 (6.3)	72.5 (4.9)	73.3 (5.7)	74.8 (6.7)
Education (years)	15.4 (2.9)	16.7 (2.1)	16.1 (2.6)	15.8 (2.6)
Gender (*n*)				
Males	10	20	30	45
Females	45	32	77	155
Race (*n*)				
Non-White/Non-Caucasian	1	0	1	1
White, Non-Hispanic	54	52	106	199
Test Interval (days)	7.6 (1.9)	6.8 (0.8)	7.3 (1.5)	7.4 (1.6)
WRAT Premorbid Intellect (*SS*)	108.6 (8.0)	110.6 (7.4)	109.6 (7.7)	108.2 (7.4)
RBANS Indexes (*SS*)				
Immediate Memory	114.0 (13.2)	107.5 (13.4)	110.8 (13.6)	108.1 (14.3)
Visuospatial/ Constructional	101.1 (12.9)	106.7 (13.8)	103.9 (13.6)	105.6 (14.2)
Language	106.0 (10.8)	104.4 (11.5)	105.2 (11.1)	104.5 (11.1)
Attention	105.5 (13.2)	108.8 (13.3)	107.1 (13.3)	105.6 (14.6)
Delayed Memory	110.6 (8.1)	110.6 (12.1)	110.6 (10.2)	108.6 (10.3)
Total Scale	110.7 (12.1)	110.9 (13.4)	110.8 (12.7)	109.2 (12.8)

Note: WRAT Premorbid Intellect = Wide Range Achievement Test Reading Subtest, RBANS = Repeatable Battery for the Assessment of Neuropsychological Status, *SS* = *Standard Score*. Combined Development Sample was composed of the 107 participants from the Combined Validation Sample, along with the 93 cognitively intact participants from Duff’s (2014) development sample.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2.

Regression equations for predicting Time 2 scores from Duff's (2014) Development sample.

	Predicted T_2	R^2	SE_{est}
Hopkins Verbal Learning Test – Revised			
Total Recall	9.18 + (T_1*0.79)	.58	3.71
Delayed Recall	8.87 + (T_1*0.50) – (*age**0.04)	.50	1.88
Brief Visuospatial Memory Test – Revised			
Total Recall	14.13 + (T_1*0.92) + (*ed**0.43) – (*age**0.14)	.66	5.29
Delayed Recall	0.28 + (T_1*0.72) + (*ed**0.18) + (*sex**0.99)	.66	1.88
Symbol Digit Modality Test	24.55 + (T_1*0.86) – (*age**0.21)	.76	4.97
Trail Making Test			
Part A	−15.10 + (T_1*0.72) + (*age**0.29)	.63	9.93
Part B	−85.54 + (T_1*0.57) + (*age**1.56)	.57	33.13

Note. All scores are raw scores. Age and education (*ed*) are in years, sex is coded as male = 0 and female = 1. R^2 = squared value of Pearson's correlation coefficient for initial and retest score, SE_{est} = Standard error of the estimate. To calculate the Predicted Time 2 (T_2) score, use the formula in the column titled "Predicted T_2". To calculate the reliable change score, use (Observed T_2 – Predicted T_2) / SE_{est}.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Table 3.

Baseline, Observed and Predicted One-Week cognitive scores, standardized *z* scores, and *p* values for difference from expectation ($z = 0$) based on the normal distribution of *z* scores in intact participants

	Observed Baseline	Observed One-Week	Predicted One-Week	z score	*p* value
Hopkins Verbal Learning Test – Revised					
Total Recall	27.53 (4.5)	30.64 (4.6)	30.93 (3.5)	−0.08 (0.9)	0.38
Delayed Recall	9.80 (1.9)	10.67 (1.6)	10.84 (1.0)	−0.09 (0.6)	0.14
Brief Visuospatial Memory Test – Revised					
Total Recall	21.47 (6.1)	29.37 (4.5)	30.52 (6.3)	−0.22 (0.8)	0.006
Delayed Recall	8.90 (2.1)	10.27 (1.4)	10.29 (1.7)	−0.01 (0.8)	0.91
Symbol Digit Modality Test	44.49 (8.0)	47.33 (9.4)	47.41 (7.4)	−0.02 (1.0)	0.87
Trail Making Test					
Part A	36.48 (13.7)	35.17 (20.2)	32.42 (10.6)	0.28 (1.8)	0.12
Part B	85.88 (39.5)	76.12 (31.6)	77.74 (26.6)	−0.05 (0.7)	0.46

Note: *p* value = significance of one-sample *t* tests examining whether *z* scores differed from expectation ($z = 0$) based on the normal distribution of *z* scores.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4.

Percentage of intact sample that displayed smaller-than-expected practice effects, expected practice effects, or greater-than-expected practice effects based on standardized regression-based methodology

	Practice Effect			
	Smaller-Than- Expected	**Expected**	**Greater-Than-Expected**	***p* value**
Hopkins Verbal Learning Test – Revised				
Total Recall	5	93	2	.32
Delayed Recall	2	98	0	.02
Brief Visuospatial Memory Test – Revised				
Total Recall	6	94	0	.06
Delayed Recall	1	97	2	.04
Symbol Digit Modality Test	4	91	5	.83
Trail Making Test				
Part A	2	94	4	.26
Part B	2	95	3	.18

Note. *p* value = significance of chi square tests between Observed distribution and Expected distribution based on the normal curve distribution of *z* scores (5% display smaller-than-expected practice effects, 90% display expected practice effects, 5% display greater-than-expected practice effects).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5.

Regression equations for predicting Time 2 scores from the current cognitively intact combined development sample

	***F* (*df*)**	***R^2***	***SE_{est} ^{a}***	***C^b***	***B^c***	**Other variables ^{d}**
Hopkins Verbal Learning Test – Revised						
∃Total Recall	151.46 (1, 198)	.43	3.34	13.13	0.65	
∃Delayed Recall	162.54 (1, 198)	.45	1.17	5.90	0.49	
Brief Visuospatial Memory Test – Revised						
∃Total Recall	49.33 (5, 194)	.56	3.78	11.01	0.52	− (age *0.13) + (education *0.17) + (sex *0.41) + (*WRAT* *0.13)
∃Delayed Recall	46.10 (5, 194)	.54	1.35	2.78	0.49	− (age *0.04) + (education *0.07) + (sex *0.23) + (*WRAT* *0.04)
Symbol Digit Modality Test	168.19 (3, 195)	.72	4.92	19.83	0.83	− (age *0.21) + (education *0.37)
Trail Making Test						
Part A	64.13 (1, 198)	.25	14.23	12.21	0.61	
Part B	45.50 (6, 193)	.59	22.69	−9.94	0.47	+ (age *1.41) - (education *0.93) − (sex *1.94) − (*WRAT* *0.22) − (*Interval* *2.12)

Note: All F-tests are significant at $p < .001$. *df* = degrees of freedom,

[a] = standard error of the estimate,

[b] = Constant,

[c] = Unstandardized beta weight for the same Time 1 Index,

[d] = Unstandardized beta weights for other variables in the equation, WRAT = Wide Range Achievement Test – Reading subtest. Age and education are in years; sex is coded *male* = 0, *female* = 1; WRAT is in Standard Score units; and Interval is in days. To calculate the predicted Time 2 score, use the following formula: (Constant value for the Index) + Unstandardized beta weight for the Index at Time 1 * score for Index at Time 1) + (Other variables in equation as noted in the Table).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6.

Case Example

	Observed Baseline	**Observed Follow-up**	**Observed Difference**	**Predicted Follow-up from Duff’s SRBs**	**Predicted Difference from Duff’s SRBs**	***z* score from Duff’s SRBs**	**Predicted Follow-up from current SRBs**	**Predicted Difference from current SRBs**	***z* score from current SRBs**
Hopkins Verbal Learning Test Revised									
Total Recall	30	27	−3	32.88	−5.88	−1.58	32.63	−5.63	**−1.69** *
Delayed Recall	10	8	−2	10.87	−2.87	−1.53	10.80	−2.80	**−2.39** *

Note: Observed Difference = Observed Follow-up – Observed Baseline. Predicted Follow-up scores are derived from the regression formula from Duff et al. (2014) Predicted Difference = Observed Follow-up – Predicted Follow-up. z = Predicted Difference/ SE_{est}.

[*] $p < .05$.