OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 36 (2021) 339–346

Archives of CLINICAL NEUROPSYCHOLOGY

Application of Different Standard Error Estimates in Reliable Change Methods

Dustin B. Hammers* and Kevin Duff

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

*Corresponding author at: Center for Alzheimer's Care, Imaging and Research, Department of Neurology, University of Utah, 650 Komas Drive #106-A, Salt Lake City, UT 84108, USA. Tel: 801-585-3929.*E-mail address*: dustin.hammers@hsc.utah.edu (D.B. Hammers).

Editorial Decision 4 September 2019

Abstract

Objective: This study attempted to clarify the applicability of standard error (SE) terms in clinical research when examining the impact of short-term practice effects on cognitive performance via reliable change methodology.
Method: This study compared McSweeney's SE of the estimate (SE_{est}) to Crawford and Howell's SE for prediction of the regression (SE_{pred}) using a developmental sample of 167 participants with either normal cognition or mild cognitive impairment (MCI) assessed twice over 1 week. One-week practice effects in older adults: Tools for assessing cognitive change. Using these SEs, previously published standardized regression-based (SRB) reliable change prediction equations were then applied to an independent sample of 143 participants with MCI.
Results: This clinical developmental sample yielded nearly identical SE values (e.g., 3.697 vs. 3.719 for HVLT-R Total Recall SE_{est} and SE_{pred}, respectively), and the resultant SRB-based discrepancy *z* scores were comparable and strongly correlated ($r = 1.0$, $p < .001$). Consequently, observed follow-up scores for our sample with MCI were consistently below expectation compared to predictions based on Duff's SRB algorithms.
Conclusions: These results appear to replicate and extend previous work showing that the calculation of the SE_{est} and SE_{pred} from a clinical sample of cognitively intact and MCI participants yields similar values and can be incorporated into SRB reliable change statistics with comparable results. As a result, neuropsychologists utilizing reliable change methods in research investigation (or clinical practice) should carefully balance mathematical accuracy and ease of use, among other factors, when determining which SE metric to use.

Keywords: Reliable change; Standard error; Assessment; Mild cognitive impairment

Introduction

Statistical procedures collectively known as reliable change methods have been developed to discriminate clinically meaningful change in serial neuropsychological assessment from repeated test exposure benefits (i.e., practice effects; Hammers, Duff, & Chelune, 2015; Lezak, Howieson, Bigler, & Tranel, 2012). Several such procedures exist, with McSweeny and colleagues' (McSweeny, Naugle, Chelune, & Luders, 1993) standardized regression-based (SRB) predicted difference method gaining wide acceptance (Attix et al., 2009; Crockford et al., 2018; Duff 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). SRB methods use linear regression to predict retest scores (Time 2) for individuals based on their baseline (Time 1) performance and other relevant information and are able to consider the impact of practice effects and other systematic biases, regression to the mean, and measurement error (Chelune, 2003; Hinton-Bayre, 2010) on repeated test performance.

doi:10.1093/arclin/acz054 Advance Access publication 15 November 2019

SRB equations generate a discrepancy change score (z score) with the difference between observed and predicted Time 2 scores in the numerator and a standard error (*SE*) term in the denominator ($z = (T_2—T_2')/SE$). The specific SE term to use has been a source of long-standing statistical debate and factors to consider include test unreliability, differential practice effects, and the inequality of variances (Hinton-Bayre, 2010; Maassen, Bossema, & Brand, 2006). Commonly utilized throughout the literature is McSweeney's (McSweeny et al., 1993) SE of the estimate (SE_{est}), which is the standard deviation of the residuals from a linear regression model. The SE_{est} is calculated from summary test statistics and is readily available on statistical software printouts. Despite McSweeney's SE_{est} prevalence throughout the literature, it has been argued that this error term fails to account for all sources of error when prediction equations are calculated in a developmental sample and are subsequently applied to an independent sample (Crawford & Howell, 1998), which is a common application of SRB equations. Specifically, although the SE_{est} accounts for uncertainty in the regression line, it ignores error arising from predicting an individual observation. In contrast, Crawford and Howell's (Crawford & Howell, 1998) SE for prediction of the regression (SE_{pred}) is a SE term that incorporates uncertainty from both the regression line and from predicting an individual test observation and is consequently calculated for each individual observation in a sample. Additionally, Crawford and Howell's approach is expected to result in more accurate standard error estimates and confidence intervals in the context of smaller sample sizes and extreme scores (Crawford & Howell, 1998). As a result of incorporating more sources of error, the SE_{pred} will be larger than the SE_{est}, which may have consequences when applying these different error terms to SRB equations to assess reliable change.

In a manuscript examining a variety of reliable change statistics and SE variables, Hinton-Bayre (2010) calculated SE_{est} and SE_{pred} values from a sample of 57 healthy control participants tested 1 year apart. The resultant error terms were then incorporated into SRB equations for a variety of commonly administered cognitive variables using a case example. The results suggested that the error terms yielded generally comparable values, with differences between the SE_{est} and SE_{pred} values being within 1% of each other (e.g., SE_{est} of 3.54 and SE_{pred} of 3.57). However, Hinton-Bayre (2010) concluded that "... while numerous authors have concluded that there is little difference in classifications across [reliable change] models in control data, there is limited systematic consideration of variations in clinical samples." As such, it can be questioned whether these different error variables would have resulted in a greater discrepancy—and subsequently different SRB discrepancy change z scores—if they were calculated under different conditions, including using clinical samples or when greater practice effects are expected. For example, Duff (2014) created regression-based prediction equations from 167 community-dwelling older adults (93 were cognitively intact and 74 were classified as having Mild Cognitive Impairment [MCI]) assessed twice over 1 week. As has been shown previously, length of test–retest interval is associated with differences in practice effects (Calamia, Markon, & Tranel, 2012), such that shorter test–retest intervals correspond to greater practice effects. Consequently, the current study sought to replicate Hinton-Bayre's previous work (2010) and extend it to this clinical sample re-tested over a shorter interval (when practice effects are likely to be their highest) by comparing the application of Duff's SRB equations for the Hopkins Verbal Learning Test—Revised (HVLT-R; Brandt & Benedict, 2001) to a validation sample when using both the SE_{est} and SE_{pred}. The aim of this research is to provide some clarity regarding the applicability of these error terms in clinical research when examining the impact of variance—including systematic bias like short-term practice effects and sources of non-systematic bias—on cognitive performance.

Method

Participants

Please see Table 1 for a description of Duff's (2014) SRB equation development sample and the current validation sample. Briefly, Duff's (2014) developmental sample, for which the SE and SRB prediction equations were calculated, included 167 community-dwelling older adults with a mean age of 78.6 (SD = 7.8) years and an average of 15.4 (SD = 2.5) years of education. The sample of participants were all Caucasian and predominantly female (81.1%). Premorbid intellect at baseline was average according to the Wide Range Achievement Test—fourth edition (WRAT-4; Wilkinson & Robertson, 2006) Reading subtest (standard score: M = 107.2, SD = 6.2), and their performances on the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, Tierney, Mohr, & Chase, 1998) were generally intact.

For the current validation sample, 143 participants were recruited from either a cognitive disorders clinic (61%) or senior centers and independent living facilities (39%). Their mean age was 75.5 (SD = 6.1, range = 65–91) years, and they averaged 16.2 (SD = 2.9, range = 12–20+) years of education. The sample of participants was evenly divided by sex (50.3% female), and the majority were Caucasian (97.9%). Premorbid intellect at baseline was average (WRAT-4 Reading subtest: M = 108.2, SD = 8.8, range = 85–145). For inclusion in the study, all participants from this sample were classified as having either single-domain or multi-domain amnestic MCI using a larger battery of cognitive tests. Classification of participants from this sample has been described previously (Duff et al., 2017). Briefly, participants were classified as amnestic MCI by participant and

Table 1. Demographic characteristics of Duff's (2014) test development and the current validation samples

	Duff (2014)	Current validation sample
Variable	Mean (*SD*)	Mean (*SD*)
n	167	143
Cognitively intact	93	0
Mild cognitive impairment	74	143
Age (years)	78.6 (7.8)	75.5 (6.1)
Education (years)	15.4 (2.5)	16.2 (2.9)
Gender (% female)	81.1%	50.3%
Race (*n*)		
African American	0	1
Hispanic/Latino American	0	1
Native American	0	1
White, non-Hispanic	167	140
Test interval (days)	7.6 (2.2)	7.2 (0.9)
WRAT-3 or -4 premorbid intellect	107.8 (6.2)	108.2 (8.8)
RBANS indexes (*SS*)		
Immediate memory	99.8 (15.4)	81.9 (16.7)
Visuospatial/constructional	106.5 (15.1)	97.7 (15.5)
Language	101.4 (11.9)	90.9 (12.3)
Attention	102.8 (14.9)	96.1 (15.2)
Delayed memory	100.1 (13.9)	77.7 (21.0)
Total scale	102.8 (13.3)	85.1 (13.2)

Note: SD = standard deviation; WRAT-3 or -4 premorbid intellect = Wide Range Achievement Test—third or fourth edition Reading Subtest; RBANS = Repeatable Battery for the Assessment of Neuropsychological Status; SS = Standard Score.

knowledgeable informant report and a baseline cognitive evaluation. Cognitive impairment for a domain was defined as a significant discrepancy (e.g., 1.5 *SD*) between current cognitive performance and an estimate of premorbid intellect. As can be observed in Table 1, on average the sample displayed below expectation abilities for immediate and delayed memory skills, particularly after considering their strong premorbid intellect, though their cognition was otherwise generally intact. 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 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 to the validation sample at a baseline visit as part of a larger battery:

- 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-min delay (range = 0–12). For all HVLT-R scores, higher raw score values indicate better performance (Durant, Duff, & Miller, 2019).
- WRAT-4 Reading (Wilkinson & Robertson, 2006) is used as an estimate of premorbid intellect, in which an individual attempts to pronounce irregular words. The score is standardized ($M = 100$, $SD = 15$) to age-matched peers, and higher values indicate better performance.
- RBANS (Randolph et al., 1998) 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 ($M = 100$, $SD = 15$), with higher scores indicating better cognition.

For the current validation sample, after approximately 1 week ($M = 7.2$ days, $SD = 0.9$, range = 6–13), the HVLT-R was repeated, with the same form being used to maximize practice effects. The RBANS and WRAT-4 were only administered at baseline.

Table 2. Complex standardized regression based change scores from Duff (2014)

Cognitive scores	Duff (2014) observed Time 1 score	Duff (2014) observed Time 2 score	Predicted T_2	$r_{xy}{}^2$	$\mathrm{SE}_{\mathrm{software}}$	$\mathrm{SE}_{\mathrm{est}}$	$\mathrm{SE}_{\mathrm{pred}}$
HVLT-R Total Recall	23.2 (5.6)	27.7 (5.7)	$9.18 + (T_1{*}0.79)$	0.582	3.71	3.697	3.719
HVLT-R delayed recall	6.7 (3.4)	9.0 (2.6)	$8.87 + (T_1{*}0.50) - (\text{age}{*}0.04)$	0.490	1.88	1.890	1.901

Note: HVLT-R = Hopkins Verbal Learning Test—Revised; T_1 = unstandardized beta weight for the Time 1 raw score; age = years old at baseline. To calculate the predicted Time 2 (T_2) score, use the formula in the column titled "Predicted T_2." $r_{xy}{}^2$ = squared value of Pearson's correlation coefficient for initial and retest score, $\mathrm{SE}_{\mathrm{software}}$ = standard error of the estimate from Duff (2014) as calculated by statistical software packages, $\mathrm{SE}_{\mathrm{est}}$ = standard error of the estimate using McSweeny et al.'s (1993) equation, $\mathrm{SE}_{\mathrm{pred}}$ = standard error for prediction of regression using Crawford and Howell's (1998) equation. To calculate the reliable change score, use either (observed T_2—predicted T_2)/$\mathrm{SE}_{\mathrm{est}}$ or (observed T_2—predicted T_2)/$\mathrm{SE}_{\mathrm{pred}}$.

Analyses

SRB group analyses. Previously published SRB prediction equations for the HVLT-R were applied to the current sample's baseline and 1-week scores. As has been described previously (Duff, 2014), the SRB prediction algorithms were calculated from a developmental sample using stepwise multiple-regression analyses to maximize the prediction of performance for each repeated measure in the cognitive battery. Specifically, the combination of demographic variables (e.g., age, education, sex), test interval, and baseline test score was used to predict the respective test score at follow-up 1 week later (Table 2).

When applying these SRB prediction equations to the current MCI sample, two different SE values were calculated for each participant. First, McSweeny and colleagues' (Hinton-Bayre, 2010; McSweeny et al., 1993) equation was used to calculate the SE of the estimate ($SE_{\mathrm{est}} = SD_y \ast \sqrt{(1—r_{xy}{}^2})$ from summary statistics, where SD_{y} = control group retest standard deviation, and r_{xy} = Pearson's correlation coefficient for initial and retest scores. Second, Crawford and Howell's (Crawford & Howell, 1998) equation was used to calculate the averaged SE for prediction of the regression ($SE_{\mathrm{pred}} = SE_{\mathrm{est}} \ast \sqrt{[1 + 1/n + (x_{\mathrm{o}}-x\bar{x})^2/SD^2{}_x(n-1)]}$) from individual test performances, where n = sample size, x_o = individual initial test score, $x\bar{x}$ = control group initial test mean, and SD_x = control group initial standard deviation. Following these calculations, z scores were then calculated for each participant's performance on the HVLT-R Total and Delayed Recall subtests, each of which reflects a normalized deviation of change for an individual participant. Specifically, the Observed One-Week score was compared to the Predicted One-Week score, normalized by both the SE_{est} (i.e., $z = (T_2—T_2')/SE_{\mathrm{est}}$) and the SE_{pred} (i.e., $z = (T_2—T_2')/SE_{\mathrm{pred}}$). Z scores calculated from both the SE_{est} and SE_{pred} for the HVLT-R Total Recall and Delayed Recall subtests were then compared to each other (e.g., SE_{est} vs. SE_{pred} for both HVLT-R subtests), as well as to expectation ($z = 0$) based on the normal distribution of z scores using a one-sample t test.

Individual distribution analyses. In a further application of the resultant z scores, they were trichotomized into "smaller-than-expected variance" (z score < -1.645), "expected variance" (z score falling between ± 1.645), or "greater-than-expected variance" (z score > 1.645) for the HVLT-R subtests. If the z scores were normally distributed, then one would expect that 5% of participants would show "smaller-than-expected variance," 90% would indicate "expected variance," and 5% would reflect "greater-than-expected variance." As alluded to previously, the term "variance" is being used to reflect changes in test performance between the two testing sessions that incorporate both systematic (practice effects) and non-systematic (measurement error, etc.) sources of bias. Using this trichotomization, individual chi-square analyses were conducted for both the HVLT-R Total Recall and Delayed subtests for z scores calculated from both the SE_{est} and SE_{pred} to determine if the observed distribution of participants deviated significantly from the expected distribution based on the normal distribution of z scores. Further, chi-square analyses were conducted comparing observed distributions of variance for HVLT-R subtests when using z scores calculated with the SE_{est} vs. the SE_{pred}.

Measures of effect size were expressed throughout as Cohen's d values for continuous data, and *Phi* coefficients for categorical data. A two-tailed alpha level was set at .05 for all statistical analyses.

Results

Differences between SE_{est} and SE_{pred}

SE $_{\mathrm{est}}$ (McSweeny et al., 1993) and SE_{pred} (Crawford & Howell, 1998) values were calculated using summary statistics and individual test performance values, respectively, from the sample that was used to create Duff's (2014) SRB prediction equations

Table 3. Baseline, observed, and predicted 1-week cognitive scores, standardized z scores using different standard error methods, and p values for difference from expectation ($z = 0$) based on the normal distribution of z scores in MCI participants

	Hopkins Verbal Learning Test—Revised	
	Total recall	Delayed recall
Observed baseline score	18.0 (5.1)	3.5 (3.3)
Observed 1-week score	21.6 (6.5)	5.8 (3.7)
Predicted 1-week score	23.4 (4.0)	7.6 (1.7)
z score using $\mathrm{SE}_{\mathrm{est}}$	−0.485 (1.1)	−0.931 (1.3)
z score using $\mathrm{SE}_{\mathrm{pred}}$	−0.482 (1.1)	−0.926 (1.3)
r between z scores using $\mathrm{SE}_{\mathrm{est}}$ and $\mathrm{SE}_{\mathrm{pred}}$	1.0	1.0
t and p values for z scores using $\mathrm{SE}_{\mathrm{est}}$ and $\mathrm{SE}_{\mathrm{pred}}$	−0.03; 0.975	−0.05; 0.960

Note: MCI = mild cognitive impairment; $\mathrm{SE}_{\mathrm{est}}$ = standard error of the estimate using McSweeny et al.'s (1993) equation; $\mathrm{SE}_{\mathrm{pred}}$ = standard error for prediction of regression using Crawford and Howell's (1998) equation, r between z scores using $\mathrm{SE}_{\mathrm{est}}$ and $S\mathrm{E}_{\mathrm{pred}}$ = Pearson product correlation coefficient between standardized regression-based z scores calculated using $\mathrm{SE}_{\mathrm{est}}$ and standardized regression-based z scores calculated using $\mathrm{SE}_{\mathrm{pred}}$, t, and p values for z scores using $\mathrm{SE}_{\mathrm{est}}$ and $\mathrm{SE}_{\mathrm{pred}}$ = paired sample t test values and significance levels between the standardized regression-based z score using $\mathrm{SE}_{\mathrm{est}}$ and the standardized regression-based z score using $\mathrm{SE}_{\mathrm{pred}}$.

for each of the HVLT-R subtests. As observed in Table 2, the SE values for the SRB equations from Duff (2014) using both methods were nearly identical (3.697 vs. 3.719, respectively, for HVLT-R Total Recall, and 1.890 vs. 1.901, respectively, for HVLT-R Delayed Recall). These values were also similar to the SE_{est} generated from statistical software. As a result, the z scores for both the HVLT-R subtests calculated using the SE_{est} and SE_{pred} were strongly correlated ($r = 1.0$, $p < .001$ for both HVLT-R Total and Delayed Recall; Table 3). When comparing the z scores for the HVLT-R subtests calculated using the SE_{est} and the z scores calculated using the SE_{pred} via paired samples t tests, no differences were observed for HVLT-R Total Recall, $t(142) = -0.03$, $p = .975$, $d = -0.005$, or HVLT-R Delayed Recall, $t(142) = -0.05$, $p = .960$, $d = -0.008$.

SRB group analyses

SRB prediction equations for each of the HVLT-R subtests from Duff (2014) were then applied to the current sample of 143 MCI participants, using both the SE_{est} and SE_{pred}. Discrepancy change scores ($z = (T_2—T_2')/SE$) for HVLT-R Total Recall and Delay Recall subtests were calculated using both SE methods and compared to expectation ($z = 0$) based on the normal distribution of z scores using one-sample t tests. Significant differences were observed when conducting these analyses for both subtests administered twice over 1 week. As a reminder, when z scores were significantly larger than zero, the current validation sample exceeded expectations based on Duff's developmental sample and reflected greater-than-expected variance over 1 week. Conversely, when z scores were significantly smaller than zero, the current validation sample fell below expectations based on Duff's developmental sample and subsequently reflected smaller-than-expected variance over 1 week. Specifically, this MCI sample displayed lower z scores than expected on HVLT-R Total Recall, $t(142) = -5.33$, $p = .001$, $d = -0.89$, and HVLT-R Delayed Recall, $t(142) = -8.42$, $p = .001$, $d = -1.41$, when using the SE_{est} to calculate the z scores, as well as HVLT-R Total Recall, $t(142) = -5.33$, $p = .001$, $d = -0.89$, and HVLT-R Delayed Recall, $t(142) = -8.42$, $p = .001$, $d = -1.41$, when using the SE_{pred} to calculate the z scores.

Individual distribution analyses

When examining the distribution of individual MCI participants that displayed "smaller-than-expected variance" (z score < -1.645), "expected variance" (z score falling between ± 1.645), or "greater-than-expected variance" (z score > 1.645) between Baseline and One-Week administrations of the HVLT-R, the majority of participants exhibited the expected level of benefit (81.3% of participants; see Table 4). However, greater proportions of individuals displayed smaller-than-expected variance over 1 week than expected based on normal distributions for both HVLT-R subtests (14% of participants for HVLT-R Total Recall, 29% of participants for HVLT-R Delayed Recall), HVLT-R Total Recall, χ^2 (2) = 19.68, $p = .001$, $Phi = 0.37$, HVLT-R Delayed Recall, χ^2 (2) = 124.21, $p = .001$, $Phi = 0.93$, using the SE_{est}, and HVLT-R Total Recall, χ^2 (2) = 19.68, $p = .001$, $Phi = 0.37$, HVLT-R Delayed Recall, χ^2 (2) = 124.21, $p = .001$, $Phi = 0.93$, using the SE_{pred}. In addition, chi-square analyses between HVLT-R subtests indicated that the distributions for each subtest were identical when using the SE_{est} and the SE_{pred} as the SE variables in the z score calculation, HVLT-R Total Recall, χ^2 (1) = 0.00, $p = .99$, $Phi = 0.00$, HVLT-R Delayed Recall, χ^2 (1) = 0.00, $p = .99$, $Phi = 0.00$. On neither subtest did greater proportions of individuals with MCI possess

Table 4. Percentage of MCI sample that displayed smaller-than-expected, expected, or greater-than-expected variance based on standardized regression-based methodology when using different standard error methods

	Variance			
	Smaller-than expected	Expected	Greater than expected	p value
HVLT-R Total Recall using $\mathrm{SE}_{\mathrm{est}}$	14.0	85.3	0.7	0.99
HVLT-R Total Recall using $\mathrm{SE}_{\mathrm{pred}}$	14.0	85.3	0.7	
HVLT-R delayed recall using $\mathrm{SE}_{\mathrm{est}}$	29.4	70.6	0.0	0.99
HVLT-R delayed recall using $\mathrm{SE}_{\mathrm{pred}}$	29.4	70.6	0.0	

Note: MCI = mild cognitive impairment; HVLT-R = Hopkins Verbal Learning Test—Revised; $\mathrm{SE}_{\mathrm{est}}$ = standard error of the estimate using McSweeny et al.'s (1993) equation; $\mathrm{SE}_{\mathrm{pred}}$ = standard error for prediction of regression using Crawford and Howell's (1998) equation; p value = significance of chi square tests comparing observed distributions of smaller-than-expected, expected, and greater-than-expected variance for HVLT-R subtests when using z scores calculated with the $\mathrm{SE}_{\mathrm{est}}$ versus $\mathrm{SE}_{\mathrm{pred}}$.

greater-than-expected variance over 1 week than anticipated based on the normal distribution of z scores (greater-than-expected variance was generally around the expected 5% value for each measure).

Discussion

The current study sought to consider how two different measurements of standard error—the standard error of the estimate of the regression (SE_{est}; McSweeny et al., 1993) and standard error for prediction of the regression (SE_{pred}; Crawford & Howell, 1998)—affected the calculation of reliable change. Specifically, in this study SE_{est} and SE_{pred} values were calculated from previously published SRB-predicted difference equations (Duff, 2014) for the HVLT-R from a developmental sample of both cognitively intact and amnestic MCI community-dwelling older adults assessed twice over a 1-week period. These resultant SE values were then incorporated into Duff's (2014) SRB equations for validation purposes using a separate and independent sample of participants with MCI.

When calculating the SE_{est} and SE_{pred} values from Duff's (2014) clinical sample over a short re-test interval, it was observed that little difference was evident between the two calculations. Specifically, the values of SE_{est} and SE_{pred} were consistently within 0.59% of each other, and the Pearson product correlation of 1.0 for the resultant SRB-based z scores suggests that the difference between these two values represented a linear transformation. The SE_{pred} was consistently slightly larger than the SE_{est}, which would be expected given that the SE_{pred} accounts for multiple sources of error (both from the regression model and for the individual observation) relative to the SE_{est} (Crawford & Howell, 1998). These results were consistent with Hinton-Bayre's previous work (2010) with a smaller sample size of healthy controls, assessed over a 1-year test–retest interval. As our sample was relatively large in size and reflected independent (non-demented) community-dwelling older adults, our study possessed favorable conditions (large sample size and non-extreme scores) described by Crawford and Howell (Crawford & Howell, 1998) for convergence in accuracy between SE calculations. These findings also appear to suggest that although the characteristics of the SRB developmental sample—including clinical sample, test–retest interval, and degree of expected practice effect—may have an effect of the overall size on the SE calculation (e.g., SE_{est} of 3.54 for HVLT-R Total Recall for Hinton-Bayre (2010) vs. SE_{est} of 3.697 for HVLT-R Total Recall currently), they have a negligible effect on the relative difference between SE_{est} and SE_{pred} calculations.

Given the near equivalence of the SE calculations, it is therefore not surprising that the resultant discrepancy z scores for the HVLT-R subtests from Duff's (2014) SRB prediction equations were comparable when applied to a validation sample ($p = .960$–$.975$). Specifically, our study observed that although the current MCI validation sample performed better at One-Week versus Baseline for both HVLT-R subtests, Observed One-Week performance fell below expectations relative to Predicted One-Week performance based on Duff's developmental sample (Cohen's $d = |0.89$–$1.41|$), and subsequently reflected smaller-than-expected variance over 1 week when using either SE calculation. Additionally, no differences were observed between the distributions of participants that displayed smaller-than-expected, expected, or greater-than-expected variance in our MCI sample based on use of the SE_{est} versus the SE_{pred} ($p = .99$ for both HVLT-R subtests). For example, 14% and 29% of MCI participants displayed smaller-than-expected variance on HVLT-R Total and Delayed Recall, respectively, consistent with previous literature reporting an absence or a reduction of practice effects in MCI across a number of cognitive measures and retest intervals (Britt et al., 2011; Calamia et al., 2012; Cooper, Lacritz, Weiner, Rosenberg, & Cullum, 2004; Darby, Maruff, Collie, & McStephen, 2002; Duff et al., 2018; Duff et al., 2017; Schrijnemaekers, de Jager, Hogervorst, & Budge, 2006).

Our results of near mathematical equivalence when using the SE_{est} and SE_{pred} suggest that it would be understandable for some researchers (and clinicians) to have uncertainty about the proper SE term to use with SRB prediction equations. As alluded to previously, the SE_{pred} is the theoretically appropriate value for use when applying SRB prediction equations developed from one sample to an independent validation sample, given its ability to account for more sources of error (error in the regression model *plus* error in predicting a specific observation). In particular, this is an issue when the developmental sample is small and the individual's score on the predictor variable is extreme (Crawford & Howell, 1998). However, the calculation of the SE_{pred} is a more laborious than using the SE_{est}, which can be either calculated by hand using summary statistics or taken directly from statistical software (or developmental sample publications as in Duff, 2014). Although Crawford should be applauded for creating statistically sophisticated online calculators for calculating an individual SE_{pred} value from summary statistics (Crawford & Garthwaite, 2007)—thus resolving the previous issue that calculation of the SE_{pred} required the possession of the developmental sample data—these calculations still need to be conducted for each individual observation in a sample. When considering research samples of 150+ participants (or when used in clinical settings for many individual patients over a period of several years), this still represents a time-intensive endeavor. These findings appear to illustrate another context (relatively large sample sizes and an absence of extreme scores) in which the SE_{est} *and* SE_{pred} approaches seem to converge; therefore, neuropsychologists incorporating reliable change methods into research investigation (or clinical practice) should carefully balance mathematical accuracy and ease-of-use, among other factors, when determining which SE metric to use.

The current study is not without limitations. First, the only conditions examined were for participants with normal cognition or MCI, assessed by the HVLT-R twice over 1 week. As a result, consideration of other domains assessed, length of retest interval, diagnostic group, and source of recruitment (e.g., clinic vs. community; Andersen et al., 2010) will be important. Additionally, these results may not generalize to more heterogeneous participants in regards to premorbid functioning, education, and race. Despite these limitations, these results appear to have replicated and extended Hinton-Bayre's (2010) previous work showing that the calculation of the SE_{est} and SE_{pred} from a clinical sample of community-dwelling cognitively normal and MCI participants yields very similar values and can be subsequently incorporated into SRB reliable change statistics with comparable results.

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

[1] Andersen, F., Engstad, T. A., Straume, B., Viitanen, M., Halvorsen, D. S., Hykkerud, S. et al. (2010). Recruitment methods in Alzheimer's disease research: General practice versus population based screening by mail. *BMC Medical Research Methodology*, *10*, 35. doi: 10.1186/1471-2288-10-35.

[2] Attix, D. K., Story, T. J., Chelune, G. J., Ball, J. D., Stutts, M. L., Hart, R. P. et al. (2009). The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*, *23*(*1*), 21–38. doi: 10.1080/13854040801945078.

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

[4] Britt, W. G., 3rd, Hansen, A. M., Bhaskerrao, S., Larsen, J. P., Petersen, F., Dickson, A. et al. (2011). Mild cognitive impairment: Prodromal Alzheimer's disease or something else? *Journal of Alzheimer's Disease*, *27*(*3*), 543–551. doi: 10.3233/JAD-2011-110740.

[5] Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(*4*), 543–570.

[6] Chelune, G. (2003). Assessing reliable neuropsychological change. In F. R (Ed.), *Prediction in forensic and neuropsychology: New approaches to psychometrically sound assessment*. Nahwah, NJ: Erlbaum.

[7] Cooper, D. B., Lacritz, L. H., Weiner, M. F., Rosenberg, R. N., & Cullum, C. M. (2004). Category fluency in mild cognitive impairment: Reduced effect of practice in test-retest conditions. *Alzheimer Disease & Associated Disorders*, *18*(*3*), 120–122.

[8] Crawford, J. R., & Garthwaite, P. H. (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, *21*(*5*), 611–620. doi: 10.1037/0894-4105.21.5.611.

[9] Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimenatal Neuropsychology*, *20*(*5*), 755–762. doi: 10.1076/jcen.20.5.755.1132.

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

[11] 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.

[12] Duff, K. (2014). One-week practice effects in older adults: Tools for assessing cognitive change. *The Clinical Neuropsychologist*, *28*(*5*), 714–725. doi: 10.1080/13854046.2014.920923.

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

[14] Duff, K., Atkinson, T. J., Suhrie, K. R., Dalley, B. C., Schaefer, S. Y., & Hammers, D. B. (2017). Short-term practice effects in mild cognitive impairment: Evaluating different methods of change. *Journal of Clinical and Experimental Neuropsychology*, *39*(*4*), 396–407. doi: 10.1080/13803395.2016.1230596.

[15] Duff, K., Beglinger, L. J., Moser, D. J., Paulsen, J. S., Schultz, S. K., & Arndt, S. (2010). Predicting cognitive change in older adults: The relative contribution of practice effects. *Archives of Clinical Neuropsychology*, *25*(*2*), 81–88.

[16] Duff, K., Schoenberg, M. R., Patton, D., Mold, J., Scott, J. G., & Adams, R. L. (2004). Predicting change with the RBANS in a community dwelling elderly sample. *Journal of the International Neuropsychological Society*, *10*(*6*), 828–834.

[17] Duff, K., Schoenberg, M. R., Patton, D., Paulsen, J. S., Bayless, J. D., Mold, J. et al. (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology*, *20*(*3*), 281–290. doi: 10.1016/j.acn.2004.07.007.

[18] Durant, J., Duff, K., & Miller, J. B. (2019). Regression-based formulas for predicting change in memory test scores in healthy older adults: Comparing use of raw versus standardized scores. *Journal of Clinical Experimental Neuropsychology*, *41*(*5*), 460–468. doi: 10.1080/13803395.2019.1571169.

[19] Gavett, B. E., Ashendorf, L., & Gurnani, A. S. (2015). Reliable change on neuropsychological tests in the uniform data set. *Journal of the International Neuropsychological Society*, *21*(*7*), 558–567. doi: 10.1017/S1355617715000582.

[20] Hammers, D., Duff, K., & Chelune, G. (2015). Assessing change of cognitive trajectories over time in later life. In Pachana, N. A., & Laidlaw, K. (Eds.), *Oxford handbook of clinical geropsychology*. Oxford: Oxford University Press.

[21] Hinton-Bayre, A. D. (2010). Deriving reliable change statistics from test-retest normative data: Comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*(*3*), 244–256. doi: 10.1093/arclin/acq008.

[22] Lezak, M., Howieson, D., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment* (5th. ed.). New York: Oxford University Press.

[23] Maassen, G. H., Bossema, E. R., & Brand, N. (2006). Reliable change assessment with practice effects in sport concussion research: A comment on Hinton-Bayre. *British Journal Sports Medicine*, *40*(*10*), 829–833. doi: 10.1136/bjsm.2005.023713.

[24] McSweeny, A., Naugle, R. I., Chelune, G. J., & 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.

[25] Randolph, C., Tierney, M. C., Mohr, E., & Chase, T. N. (1998). The repeatable battery for the assessment of neuropsychological status (RBANS): Preliminary clinical validity. *Journal of Clinical and Experimental Neuropsychology*, *20*(*3*), 310–319. doi: 10.1076/jcen.20.3.310.823.

[26] 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. *The Clinical Neuropsychologist*, *24*(*8*), 1339–1354. doi: 10.1080/13854046.2010.521770.

[27] Sanchez-Benavides, G., Pena-Casanova, J., Casals-Coll, M., Gramunt, N., Manero, R. M., Puig-Pijoan, A. et al. (2016). One-year reference norms of cognitive change in Spanish old adults: Data from the NEURONORMA sample. *Archives of Clinical Neuropsychology*, *31*(*4*), 378–388. doi: 10.1093/arclin/acw018.

[28] Schrijnemaekers, A. M., de Jager, C. A., Hogervorst, E., & Budge, M. M. (2006). Cases with mild cognitive impairment and Alzheimer's disease fail to benefit from repeated exposure to episodic memory tests as compared with controls. *Journal of Clinical and Experimental Neuropsychology*, *28*(*3*), 438–455. doi: 10.1080/13803390590935462.

[29] Stein, J., Luppa, M., Brahler, E., Konig, H. H., & Riedel-Heller, S. G. (2010). The assessment of changes in cognitive functioning: Reliable change indices for neuropsychological instruments in the elderly - a systematic review. *Dementia Geriatric Cognitive Disorders*, *29*(*3*), 275–286. doi: 10.1159/000289779.

[30] Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4: Wide Range Achievement Test, professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.