Published in final edited form as:

J Clin Exp Neuropsychol. 2019 July ; 41(5): 460–468. doi:10.1080/13803395.2019.1571169.

Regression-based formulas for predicting change in memory test scores in healthy older adults: Comparing use of raw versus standardized scores

January Durant^a, Kevin Duff^b, Justin B. Miller^a

^aNeuropsychology, Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, USA

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

Abstract

Introduction: Standardized regression based (SRB) methods can be used to determine whether meaningful changes in performance on cognitive assessments occur over time. Both raw and standardized scores have been used in SRB models but it is unclear which score metric is most appropriate for predicting follow-up performance. The aim of the present study was to examine differences in SRB prediction formulas using raw versus standard scores on two memory tests commonly used in assessment of older adults.

Method: The sample consisted of 135 healthy older adults who underwent baseline and 1-year follow-up neuropsychological assessment including the Hopkins Verbal Learning Test–Revised and Brief Visuospatial Memory Test–Revised. Regression models were fit to predict Time 2 scores from Time 1 scores and demographic variables. Separate models were fit using raw scores and standardized scores. Akaike's information criterion (AIC) was used to determine whether models using raw or standardized scores resulted in best fit. Pearson correlation and intraclass correlation coefficients were calculated between observed and predicted scores. Mean differences between observed and predicted scores were examined using pairwise *t* tests. To investigate whether a similar pattern of results would be evident using prediction formulas for nonmemory tests, all analyses were also conducted for nonmemory tests.

Results: All regression models were significant, and R^2 values for memory test raw score models were larger than those generated by standardized score models. Memory test raw score models were also a better fit based on smaller AIC values. For nonmemory tests, raw score models did not consistently outperform standardized score models. All correlations between observed and predicted Time 2 scores were significant, and none of the predicted scores significantly differed from their respective observed score.

Conclusion: For each memory measure, raw score models outperformed standardized score models. For nonmemory tests, neither score metric model consistently outperformed the other.

CONTACT January Durant durantj@ccf.org Cleveland Clinic Lou Ruvo Center for Brain Health, 888 W Bonneville Ave, Las Vegas, NV 89106, Las Vegas, NV, USA.

Disclosure statement

No potential conflict of interest was reported by the authors.

Keywords

Memory; dementia; predicting cognition; psychometric change; reliable change

Serial assessment is commonly used in clinical neuropsychology practice (Heilbrunner, Sweet, Attix, Henry, & Hart, 2010) to examine intraindividual performance over time and detect cognitive changes. Identifying meaningful changes in cognition over time is particularly important in the diagnosis and management of neurodegenerative diseases, such as mild cognitive impairment (MCI) and Alzheimer's disease (AD), which are characterized by progressive decline. Sources of change across repeated assessments are multifactorial, and, given the natural error variance inherent in neuropsychological tests, several statistical methods have been developed to assist clinicians with interpretation of changes in scores and determining whether changes are clinically meaningful (Duff, 2012; Heilbrunner et al., 2010). Evidence-based methods for assessing reliable change in performance include the simple discrepancy score, standard deviation index, reliable change index (RCI), RCI plus practice effects, and standardized regression based (SRB) formulas (see Duff, 2012, for review).

Compared to other methods of assessing reliable change, SRB methods are thought to provide a more precise estimate of relative change in cognitive performance (Duff, 2012). The SRB method was first described by McSweeney, Naugle, Chelune, and Luders (1993) and involves using multiple regression formulas to predict follow-up scores using baseline scores and other relevant available data (i.e., demographics, time between assessments). The predicted follow-up score is compared to the observed follow-up score to determine whether meaningful change in cognitive ability has occurred over time. Consistently, the best predictor of follow-up performance is baseline or Time 1 performance (Duff et al., 2010, 2005; McSweeney et al., 1993). Existing research using stepwise regression methods has returned mixed findings regarding the relative importance of demographic variables in predicting follow-up test performance in older adults. For example, in a sample of 233 community-dwelling older adults, Duff et al. (2005) found that out of the 12 subtests of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), age contributed to eight SRBs, gender to two, education to five, and race to one. In a smaller sample of 127 community-dwelling older adults, Duff et al. (2010) found that demographic variables only contributed to two of the nine models they developed for predicting follow-up performance on a battery comprised of several commonly used neuropsychological tests. Given the apparent association between sample size and influence of demographic variables in these two studies, it stands to reason that individual differences may contribute small but meaningful effects and the number of demographic predictors reaching the threshold for inclusion in hierarchical SRB models may increase with the use of larger sample sizes.

Several studies utilizing SRB methods have used raw scores in their prediction models (Duff et al., 2010, 2005; Muslimovic, Post, Speelman, De Haan, & Schmand, 2009; Ouimet, Stewart, Collins, Schindler, & Bielajew, 2009; Portaccio et al., 2010), others have used standardized scores (Busch, Lineweaver, Ferguson, & Haut, 2015; Cysique et al., 2011; Martin et al., 2002) or a combination of raw and standardized scores (Hermann et al., 1996;

Sawrie, Chelune, Naugle, & Luders, 1996). The original McSweeney et al. (1993) study used a combination of raw scores and standard scores in their models predicting follow-up performance on the Wechsler Memory Scale–Third Edition (WMS–III) and Wechsler Adult Intelligence Scale–Revised (WAIS–R). The authors compared models for both tests using raw and standardized scores, which revealed a better model fit when raw scores were used for the WMS–III but no difference in model fit for the WAIS–R. In addition, McSweeney et al. (1993) found WAIS–R norms had adequate range to encompass their entire sample but noted a restricted range of standardized scores at the lower end of the WMS–III norms. As a result, raw scores were used in the prediction models for the WMS–III but standardized scores were used for WAIS–R models. To date, there have not been any studies systematically examining which type of scores (i.e., raw vs. standardized) are the most appropriate and effective to utilize in SRB models (Duff, 2012). Potential advantages of using standardized scores are that clinicians are familiar with interpretation and assigning meaning to this metric (McSweeney et al., 1993) and that demographic variables are often already accounted for in these corrected scores. However, standardized scores tend to compress the distribution of raw scores. Individuals with disabilities or disease are often excluded from normative sets and the exclusion of these lower performing individuals may truncate the distribution (Strauss, Sherman, & Spreen, 2006). This is particularly relevant to older adult populations given the increased prevalence of disease and other medical conditions in this age group (Strauss et al., 2006). In addition, studies examining test–retest reliability of neuropsychological measures have found that raw scores, specifically for the WAIS subtests and composite scores, are associated with higher test–retest reliability than standardized scores (Calamia, Markon, & Tranel, 2013; Lemay, Bedard, Rouleau, & Tremblay, 2004). The inclusion of raw scores in SRB models may provide a better range to encompass scores at extreme ends of the distribution and a higher degree of test–retest reliability.

The current study sought to extend SRB methodology for older adults by examining prediction of performance on two widely used tests of verbal and visuospatial memory. The primary aim was to examine differences in SRB prediction formulas for the memory tests using raw scores versus standard scores. A secondary aim was to develop SRB formulas using raw scores and *t* scores that incorporate additional demographic variables to predict 1-year follow-up performance in cognitively intact older adults on the learning and delayed recall indices of the Hopkins Verbal Learning Test–Revised (HVLT–R; Brandt & Benedict, 2001) and the Brief Visuospatial Memory Test–Revised (BVMT–R; Benedict, 1997). It was hypothesized that incorporating raw scores, rather than *t* scores, into the SRB prediction formulas would provide greater accuracy of prediction. To examine whether our findings would generalize to other cognitive domains, we also examined differences in SRB prediction formulas using raw versus standardized scores for nonmemory tests.

Method

Participants

One hundred and thirty-five older adults were recruited from independent-living facilities and community senior centers following educational talks on cognitive changes associated

with aging. The sample was primarily female (83%), with a mean age of 75.5 years ($SD = 7.4$, range = 65–96), mean education of 15.4 years ($SD = 2.7$, range = 8–20), and mean estimated premorbid intelligence based on the Word Reading subtest of the Wide Range Achievement Test, 3rd edition (WRAT–3 Reading) of 107.1 ($SD = 7.9$, range = 81–126). All participants denied history of major neurological (e.g., traumatic brain injury, stroke, dementia) or psychiatric illness (e.g., schizophrenia, bipolar disorder) or current depression (either self-report or 30-item Geriatric Depression Scale, GDS, score of >12). All participants completed a brief telephone screening (Lines, McCarroll, Lipton, & Block, 2003), which has been shown to assist in identifying late-life cognitive problems, a clinical interview, baseline neuropsychological assessment, and 1-year follow-up neuropsychological assessment. Using results from the baseline assessment, individuals were classified as cognitively intact (i.e., objective memory and nonmemory performances were at least above the 7th percentile relative to a premorbid intellectual estimate: WRAT–3 Reading). The 7th percentile is 1.5 standard deviations below the mean, which is a typical demarcation point for cognitive deficits in MCI. These individuals also denied any functional impairments (e.g., assistance needed with managing money, taking medications, driving). All data were reviewed by a neuropsychologist (K.D.). Individuals classified as having MCI or dementia were not included in these analyses. All procedures were approved by the local Institutional Review Board.

Measures

Neuropsychological battery—The neuropsychological assessment battery consisted of the GDS, WRAT–3 Reading, RBANS, Controlled Oral Word Association Test (COWAT), animal fluency, Trail Making Test (TMT) Parts A and B, Symbol Digit Modalities Test (SDMT), and two memory tests described below. Standardized scores were calculated for WRAT–3 Reading and SDMT using norms presented in the test manuals. Normative data from Tombaugh, Kozak, and Rees (1999) were used to calculate age- and education-adjusted standardized scores for animal fluency, and data from Mayo’s Older Normative Studies (MOANS; Ivnik, Malec, Smith, Tangalos, & Petersen, 1996) were used to calculate age-adjusted standardized scores for COWAT and TMT Parts A and B.

Hopkins Verbal Learning Test–Revised (HVLT–R)—The HVLT–R is a word-list learning and memory test with normative data for individuals age 16 and older. Six alternate forms are available. Examiners read aloud a list of 12 words, and then the examinee is required to state the words they recall. The task is repeated over three trials, with the sum of recalled words from the initial learning trials constituting the Total Recall (HVLT–TR) raw score. After a delay of 20–25 min, the examinee is asked to recall as many of the words as they can remember, generating a Delayed Recall (HVLT–DR) raw score. Age-adjusted t scores are calculated for all measures from the test’s manual. Primary measures of interest for the present study included the raw and age-adjusted t scores for both HVLT–TR and HVLT–DR.

Brief Visuospatial Memory Test–Revised (BVMT–R)—The BVMT–R is a visual memory test with norms for adults ages 18 through 79 years. The HVLT–R and BVMT–R have similar testing formats, are conormed, and are frequently used together to assess

auditory and visual memory in neurologic populations (Benedict & Brandt, 2001). Six equivalent alternate forms of the BVMT–R are available. Examinees are presented with a display containing six figures and are then required to draw the figures they can remember. The task consists of three learning trials with points awarded for both content and location. The sum of scores from the learning trials constitutes the Total Recall (BVMT–TR) raw score and ranges from 0–36. After a delay of 25–30 min, the examinee is asked to draw as many of the figures as they can remember and in their correct location. The score from the delayed trial constitutes the Delayed Recall (BVMT–DR) raw score. Age-adjusted *t* scores are calculated from the test's manual. Primary measures of interest for the present study included the raw and age-adjusted *t* scores for both BVMT–TR and BVMT–DR.

Data analysis

Analyses were conducted using IBM Statistics for Windows (Version 23). Simultaneous multiple regression was used to develop regression equations that predict Time 2 scores from data available at Time 1. Regression models were fit to generate separate prediction algorithms for each memory test index: HVLT–TR, HVLT–DR, BVMT–TR, and BVMT–DR. In addition, regression models were fit to generate separate prediction algorithms for the following nonmemory tests: COWAT, animal fluency, TMT Parts A and B, and SDMT. Both raw scores and standardized scores were used, for a total of eight memory test models and 10 nonmemory test models. Adjusted R^2 values and Akaike's information criterion (AIC) were used to compare raw score versus standard score models to determine best fit. AIC values were calculated for each model and the relative values compared in order to ascertain the extent to which each prediction model would likely fit when applied to new data, with the model having the smallest AIC value indicating the best fit (Burnham & Anderson, 2002). Independent variables included demographic variables of age, gender, and years of education, estimated premorbid intelligence, and Time 1 score. Age was years old at Time 1. Gender was coded as male = 0 and female = 1. Pearson correlation coefficients were calculated between observed and predicted Time 2 score. In addition, intraclass correlation coefficients (ICCs) were calculated between observed and predicted Time 2 scores based on a single-measurement, absolute-agreement, two-way mixed-effects model. Mean differences between observed and predicted scores were examined using paired-samples *t* tests.

Results

The results of the prediction of Time 2 HVLT–R and BVMT–R index scores based on Time 1 score, age, gender, education, and estimated premorbid intelligence are presented in Tables 1 and 2, respectively. All memory index model fits were significant and accounted for 31% to 40% of variance in predicted HVLT–R scores, and 41% to 51% of variance in predicted BVMT–R scores. *R*-squared values for raw score models were larger than those generated by models using standardized score predictors for all memory indices. All memory index raw score models were also a better fit than their standard score counterparts based on their smaller AIC values (Burnham & Anderson, 2002). In each memory index model, Time 1 scores were a significant predictor, though in most models, demographic predictors were not significant. The exception to this was age in predicting raw HVLT–TR scores, though after accounting for multiple comparisons this would no longer be significant.

The results of the prediction of Time 2 nonmemory test scores (COWAT, animal fluency, TMT Parts A and B, and SDMT) based on Time 1 score, age, gender, education, and estimated premorbid intelligence are presented in Tables 3, 4, and 5, respectively. Nonmemory test data from two outliers were removed due to a 3-standard-deviation difference between Time 1 and Time 2 observed scores. All nonmemory test model fits were significant, and the amount of variance accounted for by the models ranged from 38% to 60%. Time 1 scores were a significant predictor for all nonmemory test models. Demographic predictors were not significant in the majority of models with the exception of age predicting animal fluency raw, TMT B raw, TMT B standardized, and SDMT standardized scores. Comparison of R -squared values indicates that raw score models accounted for a higher proportion of variance than standardized score models for animal fluency, TMT A, and TMT B. The opposite pattern was true for COWAT and SDMT, with standardized score models accounting for more variance than raw score models. Based on comparison of AIC values, raw scores were a better fit for animal fluency but standardized scores were a better fit for COWAT, TMT A, TMT B, and SDMT.

Table 6 presents performance data for each of the studied indices, the Pearson correlation, and the ICC between Time 2 observed and predicted scores. For both memory and nonmemory tests, all Pearson correlations and ICCs were moderate to strong and significant ($p < .001$). With the exception of COWAT and SDMT, correlations and ICCs were stronger for raw score models than standardized score models. Table 7 presents the range of discrepancies between the Time 2 observed and predicted scores for each memory and nonmemory test, along with the percentage of predicted scores ± 1 standard deviation above observed scores and ± 1 standard deviation below observed scores for each test. None of the predicted Time 2 scores significantly differed from their respective observed Time 2 score (i.e., all t values = 0).

Discussion

Being able to reliably determine the clinical significance of intraindividual test performance is essential, especially in neurodegenerative disease where change over time is anticipated. A number of methods exist to evaluate change over time, including regression-based prediction formulas, though it has remained unclear whether raw scores or standardized scores should be used in these models. In the present study, regression equations were established that predict 1-year follow-up performance on the HVLT-R and BVMT-R in cognitively intact older adults using baseline testing and demographic variables. Consistent with prior research, baseline performance (i.e., Time 1) was the best predictor of 1-year follow-up (i.e., Time 2) performance on the same measure, regardless of score metric used. For each memory measure, raw score models clearly outperformed standardized score models in the present samples. For nonmemory tests though, results were mixed, and neither score metric model consistently outperformed the other. In the present sample, raw scores were a better fit for animal fluency, TMT A, and TMT B based on the proportion of variance accounted for; however, standardized scores may be a better fit for COWAT, TMT A, TMT B, and SDMT if utilized in novel samples. The specific reasons for this are unclear, though this is certainly an area worthy of future study.

Demographic variables (i.e., age, education, gender, and estimate of premorbid intelligence) provided small but nonsignificant contributions to all prediction formulas, with the exception of age at Time 1, which provided a significant contribution to the HVLt-TR raw, animal fluency raw, SDMT standardized, and TMT B raw and standardized score models. While existing research examining the relative contributions of demographic variables in prediction models within an older adult population is mixed (Duff et al., 2010, 2005), it is possible that the current sample size is too small for demographic contributions to reach significance. Alternatively, it may be that while baseline performance is known to be influenced by demographic characteristics, changes in performance are not significantly related to demographic variables. Demographic contributions to performance may already be accounted for in Time 1 scores and do not contribute to prediction of Time 2 scores. The role of demographic variables in SRB prediction models is an area for further research.

Limitations of using SRB prediction models have been previously identified and are applicable to the current study (Duff, 2012; Tabachnick & Fidell, 1996). Specifically, the prediction models may be less accurate when baseline scores are toward the extreme range of the distribution, and caution must be exercised when using the models with individuals whose characteristics deviate from those of the current sample (i.e., primarily female, Caucasian, older adults, average premorbid intelligence, and cognitively intact). Future research will be needed to replicate the current findings in a larger sample and to determine the stability and utility of the HVLt-R and BVMT-R SRB prediction formulas in clinical samples (e.g., individuals referred for evaluation of memory concerns).

Whenever a study seeks to examine a sample of cognitively “normal” individuals, like the current study does, it runs the risk of presenting a biased sample, where normal variations in cognition have been removed. Such studies may not be as generalizable to the larger population. Despite these limitations, the current study presents SRB formulas that have the potential to be useful in determining whether changes between baseline and 1-year follow-up performance on the HVLt-R and BVMT-R are within the expected range for cognitively intact older adults. These formulas take into account demographic variables and baseline test performance in predicting follow-up performance on the learning and delayed recall indices of two commonly used memory tests. Using the HVLt-R and BVMT-R prediction formulas generated with raw scores, as opposed to *t* scores, may provide the most precise prediction of follow-up scores. Given our mixed findings using nonmemory test data, additional research will be needed to determine whether raw or standardized scores are more precise in nonmemory test prediction models.

Acknowledgments

Funding

This work was supported by the National Institutes on Aging [grant number 5R01AG045163]; and an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health [grant number 5P20GM109025].

References

- Benedict RHB (1997). Brief visuospatial memory test – Revised: Professional manual. Lutz, FL: Psychological Assessment Resources Inc.
- Benedict RHB, & Brandt J (2001). Hopkins verbal learning test-revised/brief visuospatial memory test-revised: Professional manual supplement. Lutz, FL: Psychological Assessment Resources Inc.
- Brandt J, & Benedict RHB (2001). Hopkins verbal learning test – Revised: Professional manual. Lutz, FL: Psychological Assessment Resources Inc.
- Burnham KP, & Anderson DR (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.) New York, NY: Springer.
- Busch RM, Lineweaver TT, Ferguson L, & Haut JS (2015). Reliable change indices and standardized regression-based change score norms for evaluating neuropsychological change in children with epilepsy. *Epilepsy & Behavior*, 47, 45–54. [PubMed: 26043163]
- Calamia M, Markon K, & Tranel D (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *The Clinical Neuropsychologist*, 27(7), 1077–1105. [PubMed: 24016131]
- Cysique LA, Franklin D Jr, Abramson I, Ellis RJ, Letendre S, Collier A, ... Heaton RK CHARTER group, & the HNRC Group. (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology*, 33(5), 505–522. [PubMed: 21391011]
- Duff K (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, 27(3), 248–261. [PubMed: 22382384]
- 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. *Archives of Clinical Neuropsychology*, 25(2), 81–88. [PubMed: 20064816]
- Duff K, Schoenberg MR, Patton D, Paulsen JS, Bayless D, Mold J, & Adams RL (2005). Regression-based formulas for predicting change in RBANS subtests with older adults. *Archives of Clinical Neuropsychology*, 20 (3), 281–290. [PubMed: 15797165]
- Heilbrunner RL, Sweet JJ, Attix DK, Henry GK, & Hart RP (2010). Official position of the American academy of clinical neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, 8, 1267–1278.
- Hermann BP, Seidenberg M, Schoenfeld J, Peterson J, Leveroni C, & Wyler AR (1996). Empirical techniques for determining the reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. *Epilepsia*, 37(10), 942–950. [PubMed: 8822692]
- Ivnik RJ, Malec JF, Smith GE, Tangalos EG, & Petersen RC (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, 10(3), 262–278.
- Lemay S, Bedard MA, Rouleau I, & Tremblay PLG (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist*, 18(2), 284–302. [PubMed: 15587675]
- Lines CR, McCarroll KA, Lipton RB, & Block GA (2003). Telephone screening for amnesic mild cognitive impairment. *Neurology*, 60(2), 261–266. [PubMed: 12552041]
- Martin R, Sawrie S, Gilliam F, Mackey M, Faught E, Knowlton R, & Kuzniecky R (2002). Determining reliable cognitive change after epilepsy surgery: Development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. *Epilepsia*, 43(12), 1551–1558. [PubMed: 12460258]
- McSweeney AJ, 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.
- Muslimovic D, Post B, Speelman JD, De Haan RJ, & Schmand B (2009). Cognitive decline in Parkinson's disease: A prospective longitudinal study. *Journal of the International Neuropsychological Society*, 15, 426–437. [PubMed: 19402929]

- Author Manuscript
- Author Manuscript
- Author Manuscript
- Author Manuscript
- Ouimet LA, Stewart A, Collins B, Schindler D, & Bielajew C (2009). Measuring neuropsychological change following breast cancer treatment: An analysis of statistical models. *Journal of Clinical and Experimental Neuropsychology*, 31(1), 73–89. [PubMed: 18608651]
- Portaccio E, Goretti B, Zipoli V, Iudice A, Pina DD, Malentacchi GM, ... Amato MP (2010). Reliability, practice effects, and change indices for Rao's brief repeatable battery. *Multiple Sclerosis*, 16(5), 611–617. [PubMed: 20207783]
- Sawrie SM, Chelune GJ, Naugle RI, & Luders HO (1996). Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society*, 2, 556–564. [PubMed: 9375160]
- Strauss E, Sherman EMS, & Spreen O (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.) New York, NY: Oxford University Press.
- Tabachnick BG, & Fidell LS (1996). *Using multivariate statistics* (3rd ed.) New York: Harper Collins.
- Tombaugh TN, Kozak J, & Rees L (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, 14(2), 167–177. [PubMed: 14590600]

Table 1.

Regression equations for predicting Hopkins Verbal Learning Test-Revised 1-year follow-up scores using data available at Time 1.

Variable	Total recall raw			Total recall <i>t</i> score			Delayed recall raw			Delayed recall <i>t</i> score		
	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]
Constant	20.9**	6.26	—	34.38*	12.19	—	4.14	3.59	—	20.58	13.31	—
Time 1	0.60**	0.09	0.55 [0.43, 0.78]	0.53**	0.09	0.47 [0.35, 0.71]	0.68**	0.11	0.52 [0.50, 0.90]	0.62**	0.11	0.49 [0.41, 0.83]
Age	-0.10*	0.05	-0.14 [-0.21, 0.00]	-0.17	0.10	-0.13 [-0.37, 0.02]	-0.03	0.03	-0.09 [-0.10, 0.02]	-0.10	0.11	-0.07 [-0.32, 0.12]
Education	0.22	0.15	0.12 [-0.08, 0.52]	0.58	0.30	0.16 [-0.02, 1.18]	0.15	0.09	0.13 [-0.03, 0.33]	0.52	0.33	0.14 [-0.14, 1.18]
Gender	0.02	0.96	0.00 [-1.88, 1.91]	0.45	1.93	0.02 [-3.36, 4.26]	-0.08	0.57	-0.01 [-1.22, 1.04]	0.05	2.11	0.00 [-4.15, 4.24]
WRAT-3	-0.05	0.09	-0.08 [-0.16, 0.06]	-0.03	0.09	-0.02 [-0.24, 0.18]	-0.01	0.03	-0.02 [-0.07, 0.06]	-0.01	0.12	0.00 [-0.25, 0.23]
Model fit	$F(5, 129) = 17.04, p < .001; R^2 = .40$			$F(5, 129) = 12.12, p < .001; R^2 = .32$			$F(5, 128) = 14.61, p < .001; R^2 = .36$			$F(5, 128) = 11.55, p < .001; R^2 = .31$		
	AIC = 383.67			AIC = 571.00			AIC = 241.43			AIC = 591.81		

Note. Time 1 = Hopkins Verbal Learning Test-Revised test score at Time 1; *SE B* = standard error for the unstandardized beta; 95% CI = 95% confidence interval for unstandardized beta; WRAT-3 = Wide Range Achievement Test-Third Edition; AIC = Akaike information criterion. To calculate the predicted 1-year follow-up score use the following formula: (constant value for the subtest) + (unstandardized beta weight for the subtest at Time 1 $\times$ raw score for the subtest at Time 1) + (age $\times$ unstandardized beta weight for age) + (education $\times$ unstandardized beta weight for education) + (gender $\times$ unstandardized beta weight for gender) + (WRAT-3 $\times$ unstandardized beta weight for WRAT-3).

* $p < .05$.

** $p < .001$.

Table 2.

Regression equations for predicting Brief Visuospatial Memory Test–Revised 1-year follow-up scores using data available at Time 1.

Variable	Total recall raw			Total recall <i>t</i> score			Delayed recall raw			Delayed recall <i>t</i> score		
	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]
Constant	14.12	7.48	—	25.91*	13.11	—	2.25	3.14	—	16.67	13.46	—
Time 1	0.76**	0.08	0.68 [0.60, 0.91]	0.76**	0.08	0.67 [0.60, 0.91]	0.70**	0.09	0.62 [0.54, 0.87]	0.69**	0.09	0.61 [0.52, 0.86]
Age	−0.06	0.06	−0.07 [−0.20, 0.06]	−0.09	0.11	−0.06 [−0.31, 0.12]	−0.02	0.03	−0.05 [−0.07, 0.04]	−0.08	0.11	−0.05 [−0.29, 0.15]
Education	0.23	0.17	0.09 [−0.14, −0.60]	0.39	0.32	0.09 [−0.25, 1.03]	0.03	0.08	0.03 [−0.12, 0.19]	0.09	0.33	0.02 [−0.56, 0.75]
Gender	0.00	1.16	0.00 [−2.30, 2.30]	−0.28	2.02	0.00 [−4.28, 3.72]	−0.29	0.49	−0.04 [−1.26, 0.68]	−0.78	2.07	−0.03 [−4.88, 3.33]
WRAT-3	−0.06	0.08	−0.07 [−0.19, 0.06]	−0.11	0.11	−0.07 [−0.32, 0.12]	0.01	0.03	0.03 [−0.04, 0.06]	0.04	0.12	0.03 [−0.19, 0.27]
Model fit	$F(5, 129) = 26.91, p < .001; R^2 = .51$			$F(5, 129) = 24.25, p < .001; R^2 = .48$			$F(5, 129) = 20.20, p < .001; R^2 = .44$			$F(5, 129) = 17.73, p < .001; R^2 = .41$		
	AIC = 439.15			AIC = 588.69			AIC = 206.01			AIC = 594.94		

Note. Time 1 = Brief Visuospatial Memory Test–Revised test score at Time 1; *SE B* = standard error for the unstandardized beta; 95% CI = 95% confidence interval for unstandardized beta; WRAT-3 = Wide Range Achievement Test–Third Edition; AIC = Akaike information criterion. To calculate the predicted 1-year follow-up score use the following formula: (constant value for the subtest) + (unstandardized beta weight for the subtest at Time 1 × raw score for the subtest at Time 1) + (age × unstandardized beta weight for age) + (education × unstandardized beta weight for education) + (gender × unstandardized beta weight for gender) + (WRAT-3 × unstandardized beta weight for WRAT-3).

* $p < .05$.

** $p < .001$.

Table 3.

Regression equations for predicting Controlled Oral Word Association Test and animal fluency 1-year follow-up scores using data available at Time 1.

Variable	COWAT raw			COWAT ss			Animal fluency raw			Animal fluency t score		
	B	SE B	β [95% CI]	B	SE B	β [95% CI]	B	SE B	β [95% CI]	B	SE B	β [95% CI]
Constant	0.13	12.33	—	-0.68	2.84	—	17.54*	7.90	—	39.51*	19.32	—
Time 1	0.87**	0.08	0.69 [0.72, 1.02]	0.90**	0.08	0.72 [0.75, 1.05]	0.47**	0.09	0.49 [0.29, 0.65]	0.53**	0.09	0.56 [0.32, 0.70]
Age	-0.18	0.10	-0.10 [-0.38, 0.02]	-0.02	0.02	-0.06 [-0.07, 0.02]	-0.20**	0.07	-0.28 [-0.33, -0.07]	-0.23	0.15	-0.19 [-0.61, 0.01]
Education	0.15	0.31	0.03 [-0.47, 0.77]	0.01	0.07	0.01 [-0.14, 0.15]	0.20	0.20	0.10 [-0.19, 0.60]	0.23	0.49	0.05 [-0.74, 1.20]
Gender	-0.75	1.97	-0.02 [-4.64, 3.14]	-0.002	0.45	-0.00 [-0.90, 0.89]	-0.29	1.16	-0.02 [-2.61, 2.03]	-0.89	2.84	-0.03 [-6.54, 4.76]
WRAT-3	0.17	0.11	0.11 [-0.05, 0.40]	0.04	0.03	0.10 [-0.02, 0.09]	0.05	0.07	0.05 [-0.08, 0.18]	0.07	0.16	0.05 [-0.24, 0.38]
Model fit	$H(5, 128) = 37.06, p < .001; R^2 = .59$			$H(5, 128) = 38.10, p < .001; R^2 = .60$			$H(5, 77) = 15.18, p < .001; R^2 = .50$			$H(5, 77) = 11.78, p < .001; R^2 = .40$		
	AIC = 576.15			AIC = 182.52			AIC = 234.21			AIC = 381.75		

Note. COWAT = Controlled Oral Word Association Test; ss = scaled score; Time 1 = COWAT or animal fluency test score at Time 1; SE B = standard error for the unstandardized beta; 95% CI = 95% confidence interval for unstandardized beta; WRAT-3 = Wide Range Achievement Test-Third Edition; AIC = Akaike information criterion. To calculate the predicted 1-year follow-up score use the following formula: (constant value for the subtest) + (unstandardized beta weight for the subtest at Time 1 $\times$ raw score for the subtest at Time 1) + (age $\times$ unstandardized beta weight for age) + (education $\times$ unstandardized beta weight for education) + (gender $\times$ unstandardized beta weight for gender) + (WRAT-3 $\times$ unstandardized beta weight for WRAT-3).

* $p < .05$.

** $p < .001$.

Table 4.

Regression equations for predicting Trail Making Test Parts A and B using 1-year follow-up scores using data available at Time 1.

Variable	Trails A raw			Trails A ss			Trails B raw			Trails B ss		
	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]
Constant	-21.51	15.18	—	9.15*	3.39	—	-61.01	47.25	—	5.43	3.05	—
Time 1	0.58**	0.07	0.59 [0.44, 0.71]	0.63**	0.08	0.59 [0.48, 0.77]	0.61**	0.07	0.59 [0.48, 0.75]	0.64**	0.07	0.63 [0.50, 0.77]
Age	0.40*	0.13	0.22 [0.15, 0.66]	-0.06	0.03	-0.15[-0.11, 0.003]	1.75*	0.43	0.27 [0.90, 2.60]	-0.06*	0.03	-0.17 [-0.11, -0.01]
Education	-0.53	0.38	-0.10 [-1.29, 0.23]	0.13	0.09	0.13 [-0.04, 0.30]	1.16	1.21	0.07 [-1.23, 3.55]	-0.05	0.08	-0.05 [-0.20, 0.11]
Gender	-1.70	2.42	-0.05 [-6.47, 3.08]	0.18	0.54	0.02 [-0.89, 1.25]	1.83	7.52	0.01 [-13.06, 16.71]	-0.11	0.48	-0.02 [-1.06, 0.85]
WRAT-3	0.15	0.13	0.09 [-0.11, 0.41]	-0.03	0.03	-0.07 [-0.08, 0.03]	-0.53	0.41	-0.09 [-1.33, 0.28]	0.04	0.03	0.13 [-0.01, 0.09]
Model fit	$H(5, 128) = 23.06, p < .001; R^2 = .47$			$H(5, 128) = 15.86, p < .001; R^2 = .38$			$H(5, 127) = 32.99, p < .001; R^2 = .57$			$H(5, 127) = 21.26, p < .001; R^2 = .46$		
	AIC = 630.04			AIC = 229.22			AIC = 928.83			AIC = 197.81		

Note. Trails A = Trail Making Test Part A; Trails B = Trail Making Test Part B; ss = scaled score; Time 1 = Trail Making Test Part A or B test score at Time 1; *SE B* = standard error for the unstandardized beta; 95% CI = 95% confidence interval for unstandardized beta; WRAT-3 = Wide Range Achievement Test-Third Edition; AIC = Akaike information criterion. To calculate the predicted 1-year follow-up score use the following formula: (constant value for the subtest) + (unstandardized beta weight for the subtest at Time 1 $\times$ raw score for the subtest at Time 1) + (age $\times$ unstandardized beta weight for age) + (education $\times$ unstandardized beta weight for education) + (gender $\times$ unstandardized beta weight for gender) + (WRAT-3 $\times$ unstandardized beta weight for WRAT-3).

* $p < .05$.** $p < .001$.

Table 5.

Regression equations for predicting Symbol Digit Modalities Test 1-year follow-up scores using data available at Time 1.

Variable	SDMT raw			SDMT t score		
	<i>B</i>	<i>SE B</i>	β [95% CI]	<i>B</i>	<i>SE B</i>	β [95% CI]
Constant	20.94	11.68	—	25.13**	12.02	—
Time 1	0.81**	0.09	0.65 [0.64, 0.98]	0.82**	0.08	0.67 [0.66, 0.99]
Age	-0.21*	0.09	-0.16 [-0.40, 0.03]	-0.21*	0.09	-0.16 [-0.39, -0.04]
Education	0.24	0.26	0.06 [-0.28, 0.75]	-0.02	0.25	-0.01 [-0.52, 0.48]
Gender	2.10	1.62	0.08 [-1.10, 5.31]	1.43	1.52	0.06 [-1.58, 4.44]
WRAT-3	-0.01	0.09	-0.01 [-0.19, 0.17]	-0.01	0.09	-0.004 [-0.17, 0.16]
Model fit	$H(5, 123) = 31.72, p < .001; R^2 = .56$			$H(5, 123) = 32.69, p < .001; R^2 = .57$		
	AIC = 503.14			AIC = 487.03		

Note. SDMT = Symbol Digit Modalities Test; Time 1 = SDMT score at Time 1; *SE B* = standard error for the unstandardized beta; 95% CI = 95% confidence interval for unstandardized beta; WRAT-3 = Wide Range Achievement Test—Third Edition; AIC = Akaike information criterion. To calculate the predicted 1-year follow-up score use the following formula: (constant value for the subtest) + (unstandardized beta weight for the subtest at Time 1 $\times$ raw score for the subtest at Time 1) + (age $\times$ unstandardized beta weight for age) + (education $\times$ unstandardized beta weight for education) + (gender $\times$ unstandardized beta weight for gender) + (WRAT-3 $\times$ unstandardized beta weight for WRAT-3).

* $p < .05$.

** $p < .001$.

Table 6.
Hopkins Verbal Learning Test–Revised, Brief Visuospatial Memory Test–Revised, nonmemory test performance.

Test	Time 1 M (SD)	Time 2 Observed M (SD)	Time 2 Predicted M (SD)	ICC [95% CI]	r
<i>Hopkins Verbal Learning Test–Revised</i>					
Total Recall raw	27.1 (4.6)	27.5 (5.1)	27.5 (3.2)	.57 [.45, .68]**	.63**
Total Recall t score	56.8 (8.6)	57.8 (9.6)	57.8 (5.5)	.49 [.35, .61]**	.57**
Delayed Recall raw	9.4 (2.3)	9.1 (3.0)	9.1 (1.8)	.54 [.40, .65]**	.60**
Delayed Recall t score	54.0 (8.3)	53.5 (10.5)	53.5 (5.9)	.48 [.33, .60]**	.56**
<i>Brief Visuospatial Memory Test–Revised</i>					
Total Recall raw	18.3 (6.2)	19.4 (7.0)	19.4 (5.0)	.68 [.58, .76]**	.72**
Total Recall t score	46.0 (10.5)	48.0 (11.8)	48.0 (8.2)	.65 [.55, .74]**	.70**
Delayed Recall raw	7.9 (2.4)	7.9 (2.8)	7.9 (1.8)	.61 [.49, .71]**	.66**
Delayed Recall t score	50.2 (9.9)	50.7 (11.3)	50.7 (7.2)	.58 [.46, .68]**	.64**
<i>Nonmemory tests</i>					
COWAT raw	38.4 (10.3)	40.1 (12.9)	40.1 (9.9)	.75 [.66, .81]**	.77**
COWAT ss	11.0 (2.4)	11.4 (3.0)	11.4 (2.3)	.75 [.67, .82]**	.77**
Animal fluency raw	18.5 (5.6)	18.8 (5.4)	18.8 (3.8)	.67 [.53, .77]**	.71**
Animal fluency t score	53.2 (13.2)	54.1 (12.3)	54.3 (8.2)	.61 [.45, .73]**	.66**
TMT A raw	40.0 (13.9)	39.0 (13.9)	38.9 (9.6)	.65 [.53, .73]**	.69**
TMT A ss	10.5 (2.7)	11.0 (2.9)	11.0 (1.8)	.56 [.43, .66]**	.62**
TMT B raw	100.5 (45.5)	96.2 (47.8)	96.2 (47.8)	.72 [.63, .80]**	.75**
TMT B ss	10.8 (2.7)	11.3 (2.7)	11.3 (1.8)	.63 [.51, .72]**	.63**
SDMT raw	42.3 (8.3)	42.9 (10.2)	42.9 (7.6)	.72 [.63, .80]**	.75**
SDMT t score	51.1 (8.0)	51.6 (9.6)	51.6 (7.3)	.73 [.64, .80]**	.81**

Note. COWAT = Controlled Oral Word Association Test; TMT A = Trail Making Test Part A; TMT B = Trail Making Test Part B; SDMT = Symbol Digit Modalities Test; ss = scaled score; ICC = intraclass correlation coefficient; CI = confidence interval.

**
p < .001.

Table 7.

Discrepancy between observed Time 2 scores and predicted Time 2 scores.

Test	Minimum	Maximum	M (SD)	% 1 SD above		% 1 SD below	
				observed Time 2 score	observed Time 2 score	observed Time 2 score	observed Time 2 score
Hopkins Verbal Learning Test-Revised							
Total Recall raw	-9.6	14.0	0 (4.0)		7.4		8.9
Total Recall <i>t</i> score	-19.2	26.6	0 (8.0)		8.9		10.4
Delayed Recall raw	-5.7	9.2	0 (2.4)		9.0		3.7
Delayed Recall <i>t</i> score	19.9	33.3	0 (8.7)		11.2		5.2
Brief Visuospatial Memory Test-Revised							
Total Recall raw	-14.6	15.3	0 (4.9)		5.9		3.7
Total Recall <i>t</i> score	-24.9	26.6	0 (8.5)		5.9		5.2
Delayed Recall raw	-4.7	7.8	0 (2.1)		8.9		7.4
Delayed Recall <i>t</i> score	-23.7	33.8	0 (8.7)		9.6		6.7
Nonmemory tests							
COWAT raw	-27.3	29.8	0 (8.2)		3.7		6.0
COWAT ss	-4.8	8.7	0 (1.9)		4.5		7.5
Animal fluency raw	-9.9	9.6	0 (3.8)		6.0		8.4
Animal fluency <i>t</i> score	-23.1	22.7	0 (9.3)		9.6		9.6
TMT A raw	-48.5	24.37	0 (10.1)		3.7		7.5
TMT A ss	-4.6	7.5	0 (2.0)		9.7		9.0
TMT B raw	-141.2	68.8	0 (31.5)		3.0		6.0
TMT B ss	-4.6	7.5	0 (2.0)		6.8		9.0
SDMT raw	-14.0	43.9	0 (6.7)		3.1		3.9
SDMT <i>t</i> score	-14.6	39.7	0 (6.3)		3.1		2.3

Note. Minimum = minimum discrepancy between predicted Time 2 score and observed Time 2 score; maximum = maximum discrepancy between predicted Time 2 score and observed Time 2 score; *M* (SD) = mean and standard deviation of discrepancies between predicted Time 2 score and observed Time 2 score; % 1 SD above observed Time 2 score = percentage of predicted Time 2 scores that are greater than or equal to 1 standard deviation above the observed Time 2 score; % 1 SD below observed Time 2 score = percentage of predicted Time 2 scores that are greater than or equal to 1 standard deviation below the observed Time 2 score; COWAT = Controlled Oral Word Association Test; TMT A = Trail Making Test Part A; TMT B = Trail Making Test Part B; SDMT = Symbol Digit Modalities Test.