

UCSF

UC San Francisco Previously Published Works

Title

Long-term test-retest reliability of the California Verbal Learning Test - second edition

Permalink

<https://escholarship.org/uc/item/74v5v7n0>

Journal

The Clinical Neuropsychologist, 31(8)

ISSN

1385-4046

Authors

Alioto, Andrea G
Kramer, Joel H
Borish, Sarah
et al.

Publication Date

2017-11-17

DOI

10.1080/13854046.2017.1310300

Peer reviewed



Published in final edited form as:

Clin Neuropsychol. 2017 November ; 31(8): 1449–1458. doi:10.1080/13854046.2017.1310300.

Long-term test-retest reliability of the California Verbal Learning Test – second edition

Andrea G. Alioto, Joel H. Kramer, Sarah Borish, John Neuhaus, Rowan Saloner, Matthew Wynn, and Jessica M. Foley

Department of Neurology, University of California, San Francisco (UCSF), San Francisco, CA, USA

Abstract

Objective—Aging is often associated with declines in episodic memory. Reliable tracking of memory requires assessment instruments that are stable over time to better understand changes potentially attributable to neurodegenerative disease. While prior studies support the test–retest reliability of memory instruments over brief intervals, follow-up testing in clinical settings typically occurs at least one-year later. The present study evaluated the long-term test–retest reliability of the California Verbal Learning Test – second edition (CVLT-2), a widely used measure of episodic learning and memory.

Method—One hundred and fifty seven healthy older adults (mean age = 68.47 years; education = 17.28 years) underwent repeat assessment at an average of 1.30 years apart. Participants underwent repeat assessment using either parallel or alternate forms at follow-up. We utilized a standardized regression-based (SRB) approach to determine statistically significant changes in test scores over time.

Results—This study revealed modest 1-year test–retest correlation coefficients for the primary CVLT-2 measures (range = .57–.69). Results of SRB formulae are provided to assist clinicians with defining clinically relevant cognitive change on the CVLT-2 while controlling for confounding factors.

Conclusions—Findings from this study support repeat test administration of the CVLT-2 over longer periods, and may enhance its applicability in determining longitudinal change in memory performance.

Keywords

Episodic memory; verbal learning; test–retest reliability; aging

Introduction

Most test–retest reliability studies assess change over relatively brief intervals that range from four to six weeks. However, follow-up evaluations in clinical settings typically occur

CONTACT Andrea G. Alioto, Andrea.Alioto@ucsf.edu.

Disclosure statement

One of the co-authors of this publication (Joel H. Kramer) is a co-author of the CVLT-2.

after much longer durations (e.g. 12+ months; American Psychological Association, 1998). Few studies have evaluated the reliability of neuropsychological assessment measures over longer retest periods (e.g. Dikmen, Heaton, Grant, & Temkin, 1999; Mitrushina & Satz, 1991; Snow, Tierney, Zorzito, Fisher, & Reid, 1988), and the limited available research has suggested that memory tests may be particularly susceptible to decreases in reliability over longer intervals (Dikmen et al., 1999).

Robust test stability is particularly relevant for memory tests. Decline in episodic memory in older individuals is often a harbinger of progressive neurological disease, and the ability to discern when a drop in performance represents true decline versus statistical margin of error is critically important.

The California Verbal Learning Test – second edition (CVLT-2; Delis, Kramer, Kaplan, & Ober, 2000) is a widely used measure of verbal episodic learning and memory (Rabin, Barr, & Burton, 2005). A list of 16 items organized into four semantic categories is presented to the subject over five immediate recall trials. Free and category cued recall is tested after short and longer (20-min) delays. A yes-no recognition trial is then presented.

Prior studies have supported the reliability of the CVLT-2 over short intervals in both clinical and non-clinical populations. For example, Benedict (2005) assessed 34 participants with multiple sclerosis, each with one-week intervals between administrations. Reliability coefficients ranged from .50 to .72 for participants receiving the standard form at both baseline and follow-up, and from .54 to .72 for those who received standard then alternate form administration. In another study, a large cohort of healthy adults underwent testing with the CVLT-2 at two time points with retest intervals ranging from 7 to 29 days (Woods, Delis, Scott, Kramer, & Holdnack, 2006). Reliability coefficients for all primary measures ranged from .80 to .84 for the standard and from .61 to .73 for the alternate form groups. Greater variability was found for the process measures, with reliability coefficients falling between .19 and .83 in the standard-standard sample, and from .06 to .74 in the standard/alternate group.

The long-term test–retest reliability of the original (versus second edition) CVLT (Delis, Kramer, Kaplan, & Ober, 1987) has been established across several studies (e.g. Dangour et al., 2010; Rannikko et al., 2015; Royall, Palmer, Chiodo, & Polk, 2012), with the primary indices (e.g. sum of trials 1–5, long-delay free recall) demonstrating particularly robust temporal stability in healthy adults (e.g. Paolo, Tröster, & Ryan, 1995). However, in contrast to strong evidence for stability of the CVLT-2 over short intervals, data on the long-term reliability of this revised measure within healthy aging samples is lacking. While researchers of one study (Lundervold, Wollschläger, & Wehling, 2014) measured age and sex-related differences in longitudinal CVLT-2 performance among healthy adults, test–retest reliability was not specifically evaluated. Further, a precise method for determining reliable change in CVLT-2 scores across long-term evaluations is needed.

The present study therefore aims to evaluate the reliability of the standard and alternate forms of the CVLT-2 over an approximate one-year interval within a population of healthy older adults. In addition to assessing the long-term reliability of the CVLT-2, we used

regression-based change formulae (McSweeney, Naugle, Chelune, & Luders, 1993) to provide statistical guidelines for detecting significant change in individual CVLT-2 profiles over time and to facilitate application to clinical practice.

Method

The participants included 157 neurologically healthy older adults between the ages of 60 and 86 years who were recruited from a longitudinal study on healthy aging and cognition; as part of this longitudinal investigation, each participant was administered the CVLT-2 on at least two separate occasions. When more than one repeat assessment was administered, results from the first two administrations were selected for the purposes of this study. Each participant was evaluated during a screening visit, which entailed an informant interview, neurological examination, and cognitive testing. Inclusion criteria as ‘neurologically healthy’ were based on a Mini-Mental State Exam (MMSE) score of ≥ 26 , Clinical Dementia Rating Score (CDR) of 0, and the absence of subject or informant report of cognitive decline. Participants were excluded if they presented with a major psychiatric disorder, neurological condition affecting cognition, dementia or mild cognitive impairment (MCI), substance abuse, major systemic medical illness, current medications likely to affect central nervous system function, sensory or motor deficits that would interfere with cognitive testing, or current depression (Geriatric Depression Scale score of >15). The study was approved by the University of California San Francisco (UCSF) committee on human research, and written informed consent was obtained from all subjects before participating.

Participants were administered the CVLT-2 according to standardized instructions (Delis et al., 2000) as part of a larger battery of neuropsychological tests by a team of psychometrists who were trained by a board certified clinical neuropsychologist. Repeat testing was conducted following a minimum of 10 months. The mean test–retest interval was 406.2 days ($SD = 47.4$). Fifty-six participants underwent repeat assessment using the standard form of the CVLT-2 on both occasions (standard/standard), whereas the other 101 individuals received the standard form at baseline and the alternate form at follow-up (standard/alternate). The demographic characteristics of the study group and their test–retest interval data are displayed in Table 1.

Since prior studies suggest that the CVLT-2 process variables yield poorer reliability across even shorter time intervals (Woods et al., 2006), only the primary indices were analyzed for the purpose of this study. Pearson product-moment correlation coefficients were calculated to assess reliability. Scores were initially winsorized using a cut point of ($z = \pm 2.0$) in order to reduce the effect of possible spurious outliers. Because winsorization did not result in meaningful change in reliability coefficients, we used unwinsorized raw scores for all analyses.

A standardized regression-based (SRB) approach was employed to determine the presence of reliable change in CVLT-2 scores (McSweeney et al., 1993). Using this approach, demographic variables (age, gender, education), retest interval, and time 1 CVLT-2 index scores were regressed on time 2 CVLT-2 primary index scores within the standard/standard and standard/alternate samples. Age was entered as total years at time 1. Gender was coded

as male = 1 and female = 2. Education was coded in total number of years achieved. Test–retest interval was defined as number of days from time 1 to time 2. Test form at time 2 was coded as standard form = 1 and alternate form = 2.

Using baseline test scores and modifiers, the SRB change score equation predicts retest performance at time 2. The difference between the predicted score and the actual score is then transformed into a *z*-score by dividing it by the standard error of the estimate (SEE). Scores that exceed the ± 1.645 , falling at the 90% confidence interval, represent clinically meaningful changes.

Results

CVLT-2 primary measure means, standard deviations, and reliability coefficients are displayed in Table 2. Reliability coefficients for the primary CVLT-2 measures ranged from .57 (recognition discriminability) to .69 (long-delay free recall).

Of note, three participants were determined to have a CDR of .5 at follow-up, suggesting possible conversion to MCI. However, reliability coefficients were comparable to the overall sample when these patients were omitted from the analysis, suggesting minimal influence of any mild cognitive changes to the results of this study. Therefore, these participants remained in the study sample, and results were unlikely to be influenced by their inclusion.

Results of regression analyses are provided in Table 3. Specifically, the coefficient of determination, SEE, and constant are provided for each CVLT-2 primary index along with beta weights for the time 1 score, test form, and relevant demographic variables. Base rates of improvement, declines, and stability of the primary CVLT-2 variables were determined. Difference scores falling in the predetermined CI were categorized as reflecting normal variability in performance, and were classified as ‘stable,’ while scores that fell outside the confidence interval were designated as having either improved or declined. Results indicated that the scores of 0–6% of participants improved, 0–8% of participants declined, and 88–97% demonstrated stability in performance over time, which is consistent with expectation given the 90% confidence interval applied.

Discussion

The present study evaluated the long-term reliability of the primary measures of the CVLT-2 over a clinically relevant (approximately one year) interval. Results revealed retest coefficients ranging from .57 to .69. Data from this study provide psychometric support for repeat test administration over longer periods, and suggests relative long-term stability of this measure among healthy older adults, which is particularly noteworthy given the protracted and clinically relevant test–retest interval.

In evaluating the size of reliability coefficients, there is considerable variability in interpreting what constitutes acceptable reliability for neuropsychological instruments. While some investigators have argued that values of at least .7 are ‘minimally acceptable’ (Cicchetti, 2001; Satler, 2001; Strauss, Sherman, & Spreen, 2006), others suggest lower cut-offs of .4 to .6 (Altman, 1991; Weintraub et al., 2014). Furthermore, Charter and Feldt

(2001) recognize the importance of considering reliability in clinical evaluation, but argue against specific cutoffs for reliability coefficients. The reliability data presented in the current study may be particularly applicable to standard clinical use, as our one-year test–retest interval more closely approximates the time frame in which possible memory changes are typically monitored in many clinical settings. Accordingly, the SRB methodology outlined in this study provides a means by which clinicians can more readily determine a statistically unusual change in these standard CVLT-2 measures.

Furthermore, these results offer an important extension of prior research that documents the adequate reliability of the CVLT-2 over intervals of less than 90-days (Benedict, 2005; Woods et al., 2006). Findings are also consistent with the small literature on long-term stability of other episodic memory measures. Examination of the one-year stability of the RAVLT, for example, found retest coefficients that ranged from .29 to .68 using parallel forms (Mitrushina & Satz, 1991; Uchiyama et al., 1995), and from .39 to .80 using alternate forms (Uchiyama et al., 1995). In another study, Woods et al. (2005) evaluated the reliability of the standard and component process measures of the HVLt-R after one-year, and found reliability coefficients ranging from .14 to .49.

The modest temporal stability found in this study may relate to several factors including the time interval employed. Most test manuals report test–retest correlations across relatively brief time intervals (e.g. days to weeks) that are far shorter than occurs during most clinical retesting scenarios (e.g. months to years), and the magnitude of test–retest correlations has been shown to diminish with increasing time periods (Duff, 2012). Furthermore, patient characteristics such as fatigue and motivation (Attix et al., 2009), as well as increased error variance related to situational factors of each testing session, can reduce the magnitude of test–retest reliability correlations over time (Heilbronner et al., 2010). The effects of age, education, and gender can also influence the degree of change across serial assessments (Heilbronner et al., 2010). For example, younger adults have shown higher retest correlations than older adults on some tests (Duff, 2012; Tombaugh, 2005). Modest correlations for intact individuals can also be due to regression to the mean in that high scores tend to decline while low scores tend to improve on subsequent testing (Strauss et al., 2006). In addition, memory measures in particular have been found to produce poorer test–retest reliabilities versus other cognitive tests (Dikmen et al., 1999); it has been suggested that lower reliability estimates may not be entirely related to methodological problems, but rather to the variable nature of this particular cognitive ability in healthy adults. Further, lower reliability may in part be due to the variable effects of practice (Dikmen et al., 1999); this may be the case even with the use of alternate forms since examinees may acquire a test-taking response strategy or may become more familiar with the testing procedure (Heilbronner et al., 2010). Greater practice effects may be particularly present in higher IQ individuals since this group tends to benefit more from previous exposure (Rapport, Brines, Theisen, & Axelrod, 1997). Ceiling effects may also be salient among highly educated individuals, since their often higher time 1 score may limit the improvement in raw score achieved at time 2. There may also be an impact of number of retests on practice effect and resulting test–retest reliability (Benedict & Zgaljardic, 1998; Ivnik, Smith, & Petersen, 1995), since the most sizable practice effect tends to occur between time 1 and time 2 (as was evaluated in this study), and higher values may be possible over later repeated

evaluations. Any combination of the aforementioned factors may be implicated in the current study including our evaluation of a memory instrument, one-year retest interval, the older age of our participants, relatively high levels of education found in our sample, and possible effects of practice. Furthermore, the perhaps apparent opposing impacts of these various factors may account for some of the remaining variance observed in our T2 scores.

Further, our post hoc analyses using a retest interval of at least 1-year (with a maximal interval of four years) found stable reliability coefficients over an even longer retest period. Reliability coefficients ranged from .58 to .68 ($N = 291$). These findings are particularly salient given that age related changes in memory performance are expected over time (Davis et al., 2003; Huh, Kramer, Gazzaley, & Delis, 2006), and when considering that such changes could compromise reliability estimates when gathered over longer periods. Therefore, results of our post hoc analyses provide further support for the association between test and retest of the CVLT-2 over time. Changes outside of our reported confidence ranges should be carefully inspected and evaluated for optimal diagnostic decision-making.

In clinical practice, SRB change scores may provide statistically based guidelines for evaluating the presence of meaningful change in a patient's performance after a one-year follow-up and can be calculated using the regression variables within Table 3. Please see Appendix 1 for a detailed example of these calculations using the SRB change score formulae. Use of the SRB approach may be most appropriate for individuals who closely resemble the study sample, including healthy older adult patients who are highly educated and of Caucasian ethnicity, and whose time 1 scores are within the range of our baseline scores as presented in Table 2. Alternatively, our data may be less applicable for individuals with cognitive impairment including MCI at time 1, and for patients of younger age, lower education, and other ethnic backgrounds. SRB calculations using our regression coefficients have been programmed into a spreadsheet to enable clinicians to easily determine meaningful change in scores at longitudinal follow-up, and are available from the corresponding author upon request. Of note, SRB change scores may be most useful when supplemented with other relevant clinical data.

The present study utilizes a combined sample of participants who were administered both parallel and alternate forms at retest. Some investigators specifically argue against the use of alternate forms due to the potential introduction of additional error variance (i.e. content sampling error) as well as time sampling error associated with test-retest paradigms (Lineweaver & Chelune, 2003), and practice effects are not entirely avoided since a test taking strategy is acquired and the test format later lacks novelty (Heilbronner et al., 2010). However, other investigators advocate for the administration of alternate forms in longitudinal assessments in order to minimize the confounding effects of practice on tests of declarative memory (Benedict, 2005; Benedict & Zgaljardic, 1998). For example, Benedict (2005) evaluation of test-retest reliability of the CVLT-2 in 34 participants with multiple sclerosis revealed reliability coefficients that were only slightly higher in the alternate form group relative to the standard form group, though participants who received the alternate form at retest exhibited less practice effects. Benedict (2005) interpreted these findings to suggest that administration of the CVLT-2 alternate form at follow-up may reduce practice effects without adversely affecting reliability. In support of these existing practice

recommendations, we also observe fairly strong test–retest reliability using alternate forms at the second time point, which again highlights the clinical relevance of such an approach. Clinical decision making should be employed to weigh the relative risks versus benefits of parallel versus alternate forms during repeat testing, and our current results can be aptly applied to either selection.

We also note some potential limitations of the current study. Most notably, the external validity of the data may be influenced by the demographic characteristics of the study sample, which entailed recruitment at a university medical center setting. In particular, it is possible that reliability information derived from the primary CVLT-2 measures using a mostly Caucasian, well-educated sample of healthy older adults does not easily generalize to most clinical populations. However, the broad age range of our participants, methods employed to statistically control for relevant demographic variables, and the lack of significant differences between the standard/standard and standard/alternate samples, may improve applicability. In addition, three participants who were deemed clinically healthy at baseline were determined to have a CDR of .5 at follow-up, suggesting possible conversion MCI. However, the absence of notable changes in reliability when these participants were removed from the analyses suggests minimal influence of these participants to the results of this study.

Examination of the R^2 values from the present SRB analyses suggests that demographic variables, retest interval, and time 1 CVLT-2 index scores account for approximately 40–50% of the variance of time 2 scores. Although systematic and random error likely explain some of the remaining variance, other currently unidentified patient factors are also likely to contribute. Future research may aim to determine additional factors that may improve this percentage, including intervening variables and events of interest (i.e. surgery, medication intervention), practice effects, and intra-individual factors (e.g. brain changes associated with normal aging, baseline IQ). Future research should additionally focus on the long-term reliability of episodic memory measures such as the CVLT-2 across longer intervals (e.g. 2+ years), as well as within various clinical populations, in order to establish guidelines for determining reliable change across diagnostic and aging groups.

Acknowledgments

Funding

This work was supported by the UCSF Alzheimer’s Disease Research Center (ADRC) [grant number P50 AG023501], the Larry L. Hillblom Foundation [grant number 2014-A-004-NET], and the National Institutes of Health, National Institute on Aging (NIH-NIA) [grant number RO1 AG032289].

References

- Altman, DG. Practical statistics for medical research. London: Chapman & Hall; 1991.
- American Psychological Association. Guidelines for the evaluation of dementia and age-related cognitive decline. *American Psychologist*. 1998; 53:1298–1303. DOI: 10.1037/0003-066X.53.12.1298 [PubMed: 9872051]
- Attix DK, Story TJ, Chelune GJ, Ball JD, Stutts ML, Hart RP, et al. The prediction of change: Normative neuropsychological trajectories. *The Clinical Neuropsychologist*. 2009; 23:21–38. DOI: 10.1080/13854040801945078 [PubMed: 18720272]

- Benedict RHB. Effects of using same versus alternate-form memory tests during short-interval repeated assessments in multiple sclerosis. *Journal of the International Neuropsychological Society*. 2005; 11:727–736. DOI: 10.1017/S1355617705050782 [PubMed: 16248908]
- Benedict RHB, Zgaljardic D. Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*. 1998; 20:339–352. DOI: 10.1076/jcen.20.3.339.822 [PubMed: 9845161]
- Charter RA, Feldt LS. Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*. 2001; 23:530–537. DOI: 10.1076/jcen.23.4.530.1227 [PubMed: 11780951]
- Cicchetti DV. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*. 2001; 23:695–700. DOI: 10.1076/jcen.23.5.695.1249 [PubMed: 11778646]
- Dangour AD, Allen E, Elbourne D, Fasey N, Fletcher AE, Hardy P, ... Uauy R. Effect of 2-y n-3 long-chain polyunsaturated fatty acid supplementation on cognitive function in older people: A randomized, double-blind, controlled trial. *The American Journal of Clinical Nutrition*. 2010; 91:1725–1732. DOI: 10.3945/ajcn.2009.29121 [PubMed: 20410089]
- Davis HP, Small SA, Stern Y, Mayeux R, Feldstein SN, Keller FR. Acquisition, recall, and forgetting of verbal information in long-term memory by young, middle-aged, and elderly individuals. *Cortex*. 2003; 39:1063–1091. DOI: 10.1016/S0010-9452(08)70878-5 [PubMed: 14584567]
- Delis, D., Kramer, J., Kaplan, E., Ober, B. California Verbal Learning Test. San Antonio, TX: The Psychological Corporation; 1987.
- Delis, DC., Kramer, JH., Kaplan, E., Ober, BA. California verbal learning test – Second. San Antonio, TX: Psychological Corporation; 2000.
- Dikmen SS, Heaton RK, Grant I, Temkin NR. Test–retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*. 1999; 5:346–356. DOI: 10.1017/S1355617799544056 [PubMed: 10349297]
- Duff K. Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*. 2012; 27:248–261. DOI: 10.1093/arclin/acr120 [PubMed: 22382384]
- Heilbrunner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. 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*. 2010; 24:1267–1278. DOI: 10.1080/13854046.2010.526785 [PubMed: 21108148]
- Huh TJ, Kramer JH, Gazzaley A, Delis DC. Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society*. 2006; 12:1–7. DOI: 10.1017/S1355617706060024 [PubMed: 16433938]
- Ivnik RJ, Smith GE, Petersen RC. Longterm stability and intercorrelations of cognitive abilities in older persons. *Psychological Assessment*. 1995; 7:155–161. DOI: 10.1037/1040-3590.7.2.155
- Lineweaver, TT., Chelune, GJ. Use of the WAIS-III in the context of serial assessments: Interpreting reliable and meaningful change. In: Tulsky, DS, Saklofske, DH., Ledbetter, MF., editors. *Clinical interpretation of the WAIS-III and WMS-III*. New York, NY: Academic Press; 2003. p. 303-337.
- Lundervold AJ, Wollschläger D, Wehling E. Age and sex related changes in episodic memory function in middle aged and older adults. *Scandinavian Journal of Psychology*. 2014; 55:225–232. DOI: 10.1111/sjop.12114 [PubMed: 24601911]
- McSweeney AJ, Naugle RI, Chelune GJ, Luders H. T scores for change: An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*. 1993; 7:300–312. DOI: 10.1080/13854049308401901
- Mitrushina M, Satz P. Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology*. 1991; 47:790–801. DOI: 10.1002/1097-4679(199111)47:6<790::AID-JCLP2270470610>3.0.CO;2-C [PubMed: 1757583]
- Paolo AM, Tröster AI, Ryan JJ. Test–retest stability of the California Verbal Learning Test in older persons. *Neuropsychology*. 1995; 11:613–615. DOI: 10.1037/0894-4105.11.4.613

- Rabin LA, Barr WB, Burton LA. Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*. 2005; 20:33–65. DOI: 10.1016/j.acn.2004.02.005 [PubMed: 15620813]
- Rannikko I, Haapea M, Miettunen J, Veijola J, Murray GK, Barnett JH, ... Jääskeläinen E. Changes in verbal learning and memory in schizophrenia and non-psychotic controls in midlife: A nine-year follow-up in the Northern Finland Birth Cohort study 1966. *Psychiatry Research*. 2015; 228:671–679. DOI: 10.1016/j.psychres.2015.04.048 [PubMed: 26168931]
- Rappaport LJ, Brines DB, Theisen ME, Axelrod BN. Full scale IQ as a mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist*. 1997; 11:375–380. DOI: 10.1080/13854049708400466
- Royall DR, Palmer R, Chiodo LK, Polk MJ. Depressive symptoms predict longitudinal change in executive control but not memory. *International Journal of Geriatric Psychiatry*. 2012; 27:89–96. DOI: 10.1002/gps.2697 [PubMed: 21351296]
- Satler, JM. Assessment of children. Cognitive applications. 4. San Diego, CA: Jerome M. Sattler Publisher; 2001.
- Snow, WG., Tierney, MC., Zorzito, ML., Fisher, RH., Reid, DW. One-year test–retest reliability of selected neuropsychological tests in older adults. Paper presented to the International Neurological Society; New Orleans. 1988 Jan.
- Strauss, E., Sherman, EMS., Spreen, O. A compendium of neuropsychological tests: Administration, norms, and commentary. 3. New York, NY: Oxford University Press; 2006.
- Tombaugh TN. Test–retest reliable coefficients and 5-year change scores for the MMSE and the 3MS. *Archives of Clinical Neuropsychology*. 2005; 20:485–503. DOI: 10.1016/j.acn.2004.11.004 [PubMed: 15896562]
- Uchiyama CL, D’Elia LE, Dellinger AM, Becker JT, Selnes OA, Wesch JE, ... Miller EN. Alternate forms of the Auditory-Verbal Learning Test: Issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology*. 1995; 10:133–146. DOI: 10.1016/0887-6177(94)E0034-M [PubMed: 14589735]
- Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Slotkin J, Gershon R. The cognition battery of the NIH toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society*. 2014; 20:1–12. DOI: 10.1017/S1355617714000320 [PubMed: 24365485]
- Woods SP, Delis DC, Scott C, Kramer JH, Holdnack JA. The California Verbal Learning Test – Second edition: Test–retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology*. 2006; 21:413–420. DOI: 10.1016/j.acn.2006.06.002 [PubMed: 16843636]
- Woods SP, Scott JC, Conover E, Marcotte TD, Heaton RK, Grant I. Test–retest reliability of component process variables within the Hopkins Verbal Learning Test – Revised. *Assessment*. 2005; 12:96–100. DOI: 10.1177/1073191104270342 [PubMed: 15695747]

Appendix 1

To apply the SRB approach in clinical practice, clinicians should examine the formulae and procedures outlined in the Methods section of this paper. To illustrate, let us assume that a 75-year-old female patient with 16 years of education who receives the standard/alternate form of the CVLT-2 achieves a long delay free recall score of 12 at baseline (T1) and 9 at one-year follow-up (T2). The clinician can calculate a predicted T2 total learning score using assessment data and demographic information as follows: predicted T2 score = $(T1\text{score} \times B_{T1}) + (\text{age} \times B_{\text{age}}) + (\text{education} \times B_{\text{education}}) + (\text{gender} \times B_{\text{gender}}) + (\text{Retest Interval} \times B_{\text{RetestInterval}}) + (T2\text{test form} \times B_{T2\text{TestForm}}) + \text{constant}$. This equation is translated as $= (12 \times .64) + (75 \times -.02) + (16 \times .06) + (2 \times .70) + (365 \times .002) + (2 \times -.31) + 3.37$, yielding a predicted T2 score of 12.02. The difference between her actual T2 score and her predicted T2 score can then be calculated and translated as follows: $z\text{-score} = (\text{actual T2}$

score – predicted T2 score)/SEE = $(9 - 12.02)/2.80 = -1.08$. Since this z -score falls within the 90% confidence interval (i.e. ± 1.645), the clinician would conclude ‘no change.’

Table 1Demographic composition of the study sample ($N = 157$).

Variable	Mean	SD	Range
Test-retest interval (days)	407.22	47.47	305–540
Age (years)	68.47	7.20	60–99
Education			
<12 years	.00%		
12 years	3.82%		
13–15 years	10.19%		
16 years	26.11%		
>16 years	59.87%		
Gender (%)			
Female	41.4%		
Male	58.6%		
Ethnicity (%)			
Caucasian	91.71%		
Hispanic	3.82%		
Asian/Pacific Islander	5.09%		
African-American	.63%		

Table 2

Test-retest data for the California Verbal Learning Test – second edition ($N = 157$).

CVLT-2 variable	Time 1	Time 2	r	M diff	SD diff
Total 1–5	51.92 (10.44)	52.67 (10.19)	.64	.75	.25
Short delay free recall	11.20 (3.09)	11.69 (3.19)	.67	.49	.10
Short delay cued recall	12.27 (2.58)	12.62 (2.40)	.66	.35	.18
Long delay free recall	11.82 (2.95)	12.19 (2.88)	.69	.37	.07
Long delay cued recall	12.48 (2.70)	12.70 (2.34)	.62	.22	.36
Recognition discriminability	3.26 (.74)	3.40 (.64)	.57	.14	.10

Notes: r = Pearson correlation coefficient; CI = confidence interval; Diff = difference.

Table 3

Regression equations for predicting Time 2 California Verbal Learning Test – second edition variables ($N = 157$).

CVLT-2 variable	R^2	SEE	C^a	B^b	B^c	B^d	B^e	B^f	B^g
Total 1–5	.44	7.76	12.76	.57	-.04	.45	2.92	.003	-2.12
Short delay free recall	.47	2.37	5.59	.63	-.04	.04	.81	-.00003	-.30
Short delay cued recall	.49	1.74	5.93	.58	-.05	.08	.67	.001	-.28
Long delay free recall	.50	2.80	3.37	.64	-.02	.06	.70	.002	-.31
Long delay cued recall	.41	1.83	7.19	.50	-.02	.02	.64	-.001	-.13
Recognition discriminability	.35	.53	1.81	.47	-.005	.02	.12	-.00004	.23

Notes: SE = standard error of the estimate; Age is in years; Education is in years; Gender is coded as 1 for male and 2 for female; Retest interval is in days; Time 2 test form is coded as 1 for standard form and 2 for alternate form.

^aConstant.

^bUnstandardized b weight for time 1 index score.

^cUnstandardized b weight for age.

^dUnstandardized b weight for education.

^eUnstandardized b weight for gender.

^fUnstandardized b weight for retest interval.

^gUnstandardized b weight for time 2 test form.