Title

Reliability and Validity of a Home-Based Self-Administered Computerized Test of Learning and Memory Using Speech Recognition

Journal

Aging Neuropsychology and Cognition, 29(5)

ISSN

1382-5585

Authors

Mackin, R Scott
Rhodes, Emma
Insel, Philip S
et al.

Publication Date

2022-09-03

DOI

10.1080/13825585.2021.1927961

Peer reviewed

Published in final edited form as:

Neuropsychol Dev Cogn B Aging Neuropsychol Cogn. 2022 September ; 29(5): 867–881. doi:10.1080/13825585.2021.1927961.

Reliability and Validity of a Home-Based Self-Administered Computerized Test of Learning and Memory Using Speech Recognition

R. Scott Mackin[1,2], **Emma Rhodes**[2,6], **Philip S. Insel**[1], **Rachel Nosheny**[1,2], **Shannon Finley**[2], **Miriam Ashford**[2], **Monica R. Camacho**[2], **Diana Truran**[2], **Kenneth Mosca**[3], **Guy Seabrook**[4], **Randall Morrison**[3], **Vaibhav Narayan**[3], **Michael Weiner**[1,2,5]

[1]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco

[2]Center for Imaging of Neurodegenerative Diseases (CIND) San Francisco Veterans Affair Medical Center

[3]Janssen Research and Development LLC

[4]Johnson & Johnson Innovation

[5]Department of Radiology, University of California, San Francisco

[6]Mental Illness Research Education and Clinical Centers, Veterans Administration Medical Center, San Francisco, CA, USA

Abstract

INTRODUCTION: The objective of this study is to evaluate the reliability and validity of the ReVeRe™ word list recall test (RWLRT), which uses speech recognition, when administered remotely and unsupervised.

METHODS: Prospective cohort study. Participants included 249 cognitively intact community dwelling older adults. Measures included clinician administered neuropsychological assessments at baseline and unsupervised remotely administered tests of cognition from 6 time-points over 6 months.

RESULTS: The RWLRT showed acceptable validity. Reliability coefficients varied across time points, with poor reliability between times 1 and 2 and fair to good reliability across the remaining five testing sessions. Practice effects were observed with repeated administration as expected.

*Corresponding author: R. Scott Mackin, Ph.D., Department of Psychiatry, University of California, San Francisco, CA 94148, Scott.mackin@ucsf.edu.

Disclosures: During the past 2 years Dr. Mackin has received research support from The National Institute of Mental Health and Janssen Research and Development LLC.

Drs Seabrook, Morrison, and Narayan and Mr. Mosca are employees of Janssen Research & Development and shareholders in Johnson and Johnson.

Dr. Weiner has served on the Scientific Advisory Boards for Pfizer, BOLT International, Neurotrope Bioscience, Alzheon, Inc., Alzheimer's Therapeutic Research Institute (ATRI), Eli Lilly, U. of Penn's Neuroscience of Behavior Initiative, National Brain Research Centre (NBRC), India, Dolby Family Ventures, LP, and ADNI.

Drs Rhodes and Ashford and Philip Insel, Monica Camacho, Diana Truran-Sacrey, and Shannon Finley reported no biomedical financial interests or potential conflicts of interest.

DISCUSSION: Unsupervised computerized tests of cognition, particularly word list learning and memory tests that use speech recognition, have significant potential for large scale early detection and long term tracking of cognitive decline due to AD.

Keywords

Online cognitive tests; memory; ReVeRe; speech recognition; reliability; validity

Introduction

In recent years there has been a tremendous increase in the number of unsupervised cognitive tests administered remotely [1–3]. Given the efficiency of data collection and lack of need for staff oversight, participant travel, and other logistical barriers, unsupervised computerized measures of cognition can be obtained for large numbers of individuals much more efficiently and at significantly lower cost than traditional neuropsychological tests. Additionally, these measures are particularly well suited for repeated assessments because they can be easily administered several times over any time interval, and improvements in performance due to test familiarity, or practice effects [4], for these intervals can be quantified. However, the psychometric properties for these measures, particularly with repeated administration [5–7]is understudied which represents a significant barrier for implementation of remotely administered cognitive tests.

Most computerized tests of cognition currently available have focused on measures of speed of information processing, working memory, and sustained and divided attention for visually mediated tasks viewed on a computer screen with key stroke responses [1–3]. While several of these tests have shown promising reliability and validity and offer the potential for more precise data collection for measures of speed of information processing and task completion time [8–10], other cognitive domains have received less attention. Specifically, measures of verbal list learning and memory have largely been underrepresented because they have not been easily adapted to unsupervised computerized assessment. Verbal list learning and memory tasks are often key tests for the identification and longitudinal monitoring of neurodegenerative diseases of aging [11–13], in part due to strong associations between free recall scores and hippocampal volumes[14,15] but also because they allow for discrimination between retrieval-based and amnestic memory profiles, which can improve diagnostic accuracy and early classification of distinct neurodegenerative syndromes and etiologies[16]. Further, measures of verbal learning and memory are reliable, independent predictors of everyday functioning, health behaviors, and quality of life in older adults[17,18] . Until recently, the lack of a reliable response format for unsupervised tests of verbal list learning and memory has resulted in overreliance on computerized recognition memory paradigms, which may be less useful in discriminating between normal aging and neurodegenerative processes[19]. Speech recognition technology is ideally suited to address this issue, however, development of computerized tests utilizing speech recognition [20,21] to measure of verbal learning and memory has been slow in comparison relative to other cognitive domains. The ReVeRe™ word list recall test (RWLRT)[22], which uses speech recognition, was developed by Janssen Research & Development to address this unmet need. The RWLRT was designed to be administered remotely and without supervision using an iPad. The purpose of this

study was to evaluate the convergent and divergent validity of RWLRT with other clinician administered neuropsychological tests and to evaluate the reliability of RWLRT performance over six test administrations spanning 18 months.

Procedures

Participants for this study were referred from the BrainHealthRegistry.org (BHR). The BHR functions within the University of California, San Francisco (UCSF) and is approved by the UCSF institutional review board. BHR registrants receive no compensation for completing study procedures. Currently, more than 70,000 participants have registered with the BHR[2,3]. For this study, inclusion criteria included age greater than 60 years old, fluent in English, access to the internet and wireless internet access, and ability to provide consent. Participants were not eligible to participate if they had diagnosis or evidence of dementia, evidence of acute or uncontrolled medical illness, recent history (<6 months) of drug or alcohol abuse or dependence, or had diagnosis of significant neurological disease. Eligible participants from the BHR were referred to the study. A total of 249 participants were enrolled in the study. Once enrolled, participants completed baseline assessments in the clinic, which included traditional neuropsychological measures and the RWLRT administered on an iPad. iPads with preinstalled ReVeRe software were distributed to all study participants at the baseline visit and used for all follow-up RWLRT assessments. The RWLRT was completed by participants at five follow-up time points, for a total of six assessments: Time 1 (baseline), Time 2 (7 days post), Time 3 (21 days post), Time 4 (6 months post), Time 5 (12 months post), and Time 6 (18 months post). The RWLRT was completed without an examiner present at all time points, including the baseline in clinic visit. At baseline participants completed traditional neuropsychological tests and then completed the RWLR, and test administration was not counterbalanced. Traditional neuropsychological tests were scored using available normative data based on age, with comparable scaled scores obtained. Cognitive assessments were performed by research assistants who were supervised by a licensed clinical neuropsychologist.

Measures

Unsupervised Measures of Cognition:

Unsupervised cognitive test performance was assessed with the ReVeRe™ cognitive test platform, which uses computer software to administer and score objective cognitive tests to determine an individual's current level of cognitive functioning. This system is a custom-built iOS application developed by Janssen Research & Development and deployed on an iPad Air or later generation device. The software was built using a large variety of native libraries which are needed for functionality, such as audio capture and playback, screen layouts, and basic subject/test system interactions. The application interacts with the company repository for data storage. All data are sent from the ReVeRe™ system to the company repository via wireless internet connection using Java Script Object Notation scripts with a Secure Shell encrypted network protocol to ensure a secure connection for file transfer. Speech recognition technology was used to automate the scoring of the word list recall task, with the goal of assessing the accuracy of speech recognition technology in

the home. Prior in-clinic uses of the ReVeRe™ speech recognition engine (SRE) exceeded a 97% accuracy for the word list recall task, a metric obtained by having the speech recognition score the audio of the subject and having two independent raters score the same audio files[22]. Accuracy of home administration of the ReVeRe SRE was assessed by two independent raters who transcribed verbal responses recorded on a sample of ReVeRe Word List Recall Test audio files (n = 906). Discrepancies between responses transcribed by the two raters were resolved by consensus with a third independent rater. Final transcribed ratings were compared to the automated ReVeRe SRE output to establish overall accuracy of the SRE. Accuracy was calculated as the percent of overlap between responses detected via SRE and responses coded by human raters.

The ReVeRe™ Word List Recall (RWLRT) is based on the Rey Auditory Verbal Learning Test[23]. The memory test included a list of 15 words presented aloud, one at a time, through the iPad. After the 15-item word list is presented (Word List A), the participant is asked to verbally recall as many of the words as possible. Word List A is presented and recalled 2 more times for a total of 3 learning trials. After the final learning trial and recall is recorded, a different 15-word list (distractor list) is presented in the same manner and the subject is asked to immediately recall as many words as possible from this new list. Following the distractor list recall, the participant is asked to recall as many words as possible from Word List A. Twenty minutes after completion of the list recall tasks, the subject is asked to recall as many words as possible from Word List A. During the delay interval participants completed additional cognitive tasks on the ReVeRe platform. Verbal responses from participants were recorded and scored using voice recognition software. Performance on the learning trials of the original word list (Total Learning score) and on Delayed Recall were used as the primary scores. RWLRT stimuli (Word List A, Distractor List) were identical to AVLT stimuli, and the same version was used across all time points. Recognition memory and rates of intrusion and repetition errors were not collected for this project.

Clinician Administered Measures of Cognitive Functioning.

Verbal Learning and Memory.—Word list learning and memory was evaluated using The California Verbal Learning Test – Second Edition (CVLT-II) [24]. Learning and memory of a short story were assessed using the Logical Memory (LM) test from the Wechsler Memory Scale – Revised [25].

Information Processing Speed.—The WAIS-IV Coding and Trail Making Test Part were used to measure visuomotor speed and speed of information processing [25,26].

Working Memory.—The WAIS-IV Digit Span test was used to assess attention and sequencing abilities[25,27].

Executive Functioning.—Two measures of executive function were administered. The Controlled Oral Word Association Test (COWAT – FAS) was used to assess timed initiation, generativity, and cognitive control [28]. The Trail Making Test Part B was also used to assess sequencing and inhibition on a timed task [26].

Data Analysis

Attrition was assessed using a generalized linear mixed model with a binomial missing indicator regressed on age, sex, race, education, time, and the interaction between time and demographic variables. Test-retest reliability across all time points was evaluated using multiple intraclass correlation coefficient (ICC) estimates on a subset of the overall sample with complete data for all subtests across all time points (n = 137). Multiple ICC estimates were reported based on guidelines for appropriate selection of model, type, and definition (described below using Shrout & Fleiss terminology), [29,30] Overall consistency across time points was assessed with ICC estimates for a two-way random effects model based on absolute agreement of the average of multiple administrations (ICC 2, 6), which estimates reliability when limited to a single administration, and a two-way mixed effects model based on consistency of an average of multiple administrations (ICC 3, 6), which is appropriate for a measure that is intended to be repeated over time. Reliability between consecutive RWLRT time points was assessed using two-way random effects models based on absolute agreement of single measurements (ICC 2, 1), which treat practice effects as systematic measurement error, and two-way mixed-effects models based on consistent of single measurements (ICC 3, 1), which do not account for practice effects[29]. Reliability ICCs were interpreted as follows: <0.40, poor; 0.40–0.59, fair; 0.60–0.75, good; >0.75, excellent[31,32]. Practice effects were evaluated across adjacent time points with paired t-tests/Wilcoxon signed rank test and Cohen's *d* effect sizes[33]. Convergent validity was assessed using Pearson's correlations between gold standard neuropsychological tests (CVLT learning and delayed memory) and RWLRT subtests at the baseline evaluation. Discriminant validity was assessed using Steiger's *z* to assess for significant differences between Pearson's correlations for performance on conceptually similar tests and those for conceptually dissimilar tests.

Results

Characteristics of the sample are shown in Table 1. The sample consisted of 249 community-dwelling adults with a mean age of 70.6 (SD = 6.8) and 16.7 years of education (SD = 2.3). The sample was 48.9% female and predominantly Caucasian (81.9%) and included participants who identified as African American (2.8%), Asian (7.3%), Latino (0.8%), Pacific Islander (0.4%), and more than one race (4.4%) as well as 6 participants (2.4%) who declined to disclose racial identity. The sample was largely cognitively intact, with mean performance across gold standard neuropsychological measures in the average to high average range relative to age matched peers using published normative data. Raw score performance on traditional neuropsychological measures is shown in Table 2. Of the 249 enrolled participants, 20.1% (n = 50) were lost to follow up by the final time point, with Caucasian ($\beta = -0.84$, SE = 0.38, $p = .028$) and more highly educated ($\beta = 0.47$, SE = 0.15, $p = .002$) participants more likely to have missing data over time. Age and gender were not associated with attrition over time (all p's > .05). The accuracy of the speech recognition engine used for the in-home administration of the ReVeRe battery was 91.9%.

Practice Effects

Mean RWLRT performance across time points is displayed in Figure 1. Performance on RWLRT subtests improved significantly over time (all p-values <.05). Practice effects varied

across tasks and time points (see Table 3), with associated effect size values (*d*) ranging from 0.002 (RWLRT: Short Delay) to 1.05 (RWLRT: Delayed Recall). Magnitude of practice effects for RWLRT subtests was greatest between baseline (Time 1) and 7 days (Time 2) (mean $d = 0.89$) and smallest between 6 months (Time 4), 12 months (Time 5), and 18 months (Time 6, all *d*'s = 0.11).

Test-Retest Reliability

ICC estimates for reliability across time points are show in Table 4. ICC estimates for the overall reliability of RWLRT scores across all time points showed good consistency (ICC 3, 6) and absolute agreement (ICC 2, 6) based on the mean of multiple measurements, with estimates ranging from 0.84 (Total Learning) to 0.90 (Delayed Recall). ICC estimates for the reliability of RWLRT scores at consecutive time points (ICC 2, 1 & 3, 1) ranged from poor (Total Learning Time 1–2, ICC 2,1 = 0.35) to excellent (Long Delay Time 2–3, ICC 3,1 = 0.78). All RWLRT scores demonstrated poor to fair reliability estimates for performance at Times 1 and 2 and fair to good reliability estimates for all remaining consecutive timepoints. There were no differences in demographic or cognitive variables between the total sample and the reliability subsample (n=137, all *p*'s > .05).

Concurrent and Discriminant Validity

Correlations between baseline performance on RWLRT subtests and clinic administered neuropsychological tests of verbal learning and memory are presented in Table 4. Concurrent validity between RWLRT and a test of verbal list learning and memory (CVLT) ranged from 0.54 – 0.59. For a verbal learning and memory test for stories (LM) correlations ranged from 0.33–0.40. Evidence of discriminant validity was supported by significantly lower correlations between RWLRT subtests and clinic administered neuropsychological measures from theoretically unrelated cognitive domains (see Supplemental Table 1 for full correlation matrix).

Discussion

The results of this study indicate that the RWLRT administered remotely and unsupervised showed variable reliability across timepoints, with optimum reliability when assessing the average of multiple measurements (ICC 2,6 & 3,6) and accounting for early practice effects (ICC 3,1). Convergent and divergent validity with clinician administered cognitive assessments was also within expectation. Each of these findings will be discussed below.

The word recall task (RWLRT), which used voice recognition software, showed overall test-retest reliability lower than clinician administered versions of these test paradigms, such as the RAVLT and CVLT, which have reliability coefficients ranging from 0.67–0.9[34] and .80-.84[24], respectively, typically reported for learning trials, short delay recall, long delay recall. With repeated administrations, we report good reliability between the second and third administration (day 7 and day 21) and poor to fair reliability between the first and second time points (baseline and day 7), regardless of ICC type. Additionally, reliability was lowest when measuring absolute agreement between time points with a two-way random effects model (ICC 2,1), as opposed to measuring consistency with a

two-way mixed-effects model (ICC 3,1), Taken together, this pattern suggests improvement in performance due to previous exposure to the test, i.e., practice effect [35]. Practice effects are commonly seen with neuropsychological tests particularly when given over relatively short periods of time [36] in cognitively intact patient populations. Further, at the 6-month evaluation (Time 4), performance showed slight declines followed by more stable levels of performance across the remaining six-month intervals. As a result, reliability of the RWLRT was optimized between the second and third administration consistent with findings from other unsupervised online cognitive assessment studies [7]. These results suggest that repeated assessments over relatively short periods of time [37,38] would likely be the most accurate means of assessing cognition particularly for unsupervised online measures, which show significant practice effects over repeated administrations and greater reliability when systematic error from practice effects are excluded from ICC estimates (ICC 3,1)[39]. Unsupervised online measures may therefore be well-suited for measurement burst designs, which capitalize on practice effects by modeling individual variability in repeated administration over shorter periods of time[38]. Alternatively, development of practice trials preceding each test or alternate versions of tests may be beneficial.

All RWLRT scores demonstrated moderate convergent validity with clinician-based measures of verbal episodic memory for a similar test paradigm (CVLT). Convergent validity between primary outcome measures of the RWLRT and CVLT were notably higher than those reported for other computerized tests of verbal memory, which range from .38[40] to 0.46 [41]. Thus, the RWLRT shows an advantage over other existing computerized assessments of verbal memory, most likely due to differences in test design and response format. To our knowledge, all other computerized tests of verbal memory rely on forced choice recognition paradigms, which measure memory processes that are behaviorally and neuroanatomically dissociable from those measured by free recall paradigms[42], while the RWLRT is able to assess free recall of verbal stimuli using speech recognition software. A previous version of the RLWRT, in which participant responses were audio recorded and scored by raters, showed slightly higher convergent validity with the RAVLT, ranging from 0.63 to 0.70[22]. This suggests that agreement between RWLRT and traditional word list learning tasks is improved by allowing for more equivalence in test design and response format with some decrement potentially introduced by speech recognition accuracy. A prior investigation found >97% accuracy of RWLRT speech recognition software relative to consensus scoring of audio recordings two independent raters, while our results showed slightly lower accuracy at 91.9%[22]. This may be due to the use of naturalistic settings for test administration in our study, which carries an increased risk for extraneous noise in the testing environment. The continued use of machine learning techniques has the potential to optimize speech recognition technology to adapt to environmental confounds and reduce error.

In contrast, convergent validity between RWLRT scores and a clinician-based story learning and memory task (LM) were weaker overall, ranging from 0.33 (RWLRT: Short Delay – LM I) to 0.4 (RWLRT: Long Delay – LM II), but consistent with reported associations between other computerized tests of verbal memory and traditional story learning paradigms, which range from 0.17[41,43] to 0.52[32, 44]. Lower RWLRT validity coefficients for story learning and memory could reflect differences in task stimuli (list vs. story), number of presentations

(5 vs. 1), or targeted memory processes (i.e., recall vs. recognition). While both story and list learning paradigms are as sensitive to AD pathology, list learning tests have been found to have greater accuracy in distinguishing normal aging from MCI and higher predictive validity AD [45]. In particular, verbal free recall scores have been identified as the measure most sensitive to memory decline in older adults[46,47]. A limiting factor in comparing construct validity of the RWLRT with other computerized measures of learning and memory is the relative dearth of automated computerized verbal memory tasks. Other widely cited computerized cognitive batteries rely on visuospatial maze and pattern/face matching memory paradigms, which show relatively weak associations with traditional verbal memory tests, with *r*'s ranging from 0.14 to 0.28, but moderate to strong associations with traditional visuospatial memory tests[48–50].

In addition to moderate convergent validity with traditional verbal memory tests, the RWLRT demonstrated adequate discriminant validity when compared to traditional neuropsychological measures of attention, working memory, processing speed, language, motor speed and dexterity, and executive functioning. All RWLRT scores showed weak to very weak associations with traditional cognitive tests designed to measure non-memory domains (all *r's* <.29). Weak but statistically significant associations between RWLRT scores and measures of working memory, rapid set-shifting, processing speed, and motor speed reflect expected overlap from component cognitive processes. Of note, the RWLRT Learning score showed more significant relations with measures of processing speed and executive functioning compared to RWLRT Short and Long Delay recall scores, which is consistent with known associations among measures of fluid intelligence (i.e., executive functions, processing speed, working memory) and widely cited executive contributions to verbal list learning[51,52]. RWLRT divergent validity coefficients were consistent with those reported for other computerized cognitive assessments [41,43,49] and slightly lower than those reported for traditional clinic-based verbal list learning tests[53,54], likely reflecting a greater equivalence of testing environment among computerized assessments.

A limitation of this study is that we largely studied highly educated older adults who showed high levels of performance on clinician administered measures of cognition, including measures of verbal learning and memory. Thus, there is limited generalizability to the general population and for individuals with mild cognitive impairments. Further, recruitment of study participants from the Brain Health registry may have resulted in a sample with elevated memory concerns. An additional methodological consideration is the lack of consensus on selection and interpretation of ICCs in the measurement of test-retest reliability. Our findings were interpreted using guidelines specific to the evaluation of ICCs for standardized psychological assessments (including cognitive measures), which are notably less stringent than more general guidelines for neuropsychological measures[55,56], which do not account for variability in reliability type or coefficient. Finally, our conclusions about the suitability of the RWLRT do not account for factors that may impact the accuracy of speech recognition software, which is integral to unsupervised cognitive assessment. Future research on unsupervised or self-administered neuropsychological assessments using speech recognition technology should include analysis of speech recognition accuracy across individuals, settings, and timepoints. Despite these limitations, our results support the validity and reliability for RWLRT for use as unsupervised online screening tools and

longitudinal measures of cognition. Further, our results suggest that repeated assessments over short periods of time may yield more reliable measures of cognition which is particularly scalable for unsupervised applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

The Ray and Dagmar Dolby Family Fund

Data collection and sharing for this project was funded by Janssen Research and Development LLC & Johnson & Johnson Innovation

References

1. Cinar N, Sahiner TAH. Effects of online computerized cognitive training program Beynex on the cognitive tests of individuals with subjective cognitive impairment (SCI) and Alzheimer disease on rivastigmine therapy. Turk J Med Sci 2019.
2. Mackin RS, Insel PS, Truran D, et al. Unsupervised online neuropsychological test performance for individuals with mild cognitive impairment and dementia: Results from the Brain Health Registry. Alzheimers Dement (Amst) 2018;10:573–582. [PubMed: 30406176]
3. Weiner MW, Nosheny R, Camacho M, et al. The Brain Health Registry: An internet-based platform for recruitment, assessment, and longitudinal monitoring of participants for neuroscience studies. Alzheimers Dement 2018;14(8):1063–1076. [PubMed: 29754989]
4. White N, Forsyth B, Lee A, Machado L. Repeated computerized cognitive testing: Performance shifts and test-retest reliability in healthy young adults. Psychol Assess 2018;30(4):539–549. [PubMed: 28557476]
5. Valdes EG, Sadeq NA, Harrison Bush AL, Morgan D, Andel R. Regular cognitive self-monitoring in community-dwelling older adults using an internet-based tool. J Clin Exp Neuropsychol 2016;38(9):1026–1037. [PubMed: 27266359]
6. Wild K, Howieson D, Webbe F, Seelye A, Kaye J. Status of computerized cognitive testing in aging: A systematic review. Alzheimer's & Dementia 2008;4(6):428–437.
7. Stricker NH, Lundt ES, Edwards KK, et al. Comparison of PC and iPad administrations of the Cogstate Brief Battery in the Mayo Clinic Study of Aging: Assessing cross-modality equivalence of computerized neuropsychological tests. Clin Neuropsychol 2019;33(6):1102–1126. [PubMed: 30417735]
8. Domen AC, van de Weijer SCF, Jaspers MW, Denys D, Nieman DH. The validation of a new online cognitive assessment tool: The MyCognition Quotient. Int J Methods Psychiatr Res 2019;28(3):e1775. [PubMed: 30761648]
9. Feenstra HE, Vermeulen IE, Murre JM, Schagen SB. Online cognition: factors facilitating reliable online neuropsychological test results. Clin Neuropsychol 2017;31(1):59–84. [PubMed: 27266677]
10. Feenstra HEM, Murre JMJ, Vermeulen IE, Kieffer JM, Schagen SB. Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan. J Clin Exp Neuropsychol 2018;40(3):253–273. [PubMed: 28671504]
11. Thomas KR, Edmonds EC, Eppig J, Salmon DP, Bondi MW, Alzheimer's Disease Neuroimaging I. Using Neuropsychological Process Scores to Identify Subtle Cognitive Decline and Predict Progression to Mild Cognitive Impairment. J Alzheimers Dis 2018;64(1):195–204. [PubMed: 29865077]
12. Turchetta CS, Perri R, Fadda L, et al. Forgetting Rate on the Recency Portion of a Word List Differentiates Mild to Moderate Alzheimer's Disease from Other Forms of Dementi. J Alzheimers Dis 2018;66(2):461–470. [PubMed: 30320591]

13. Baker JE, Lim YY, Jaeger J, et al. Episodic Memory and Learning Dysfunction Over an 18-Month Period in Preclinical and Prodromal Alzheimer's Disease. J Alzheimers Dis 2018;65(3):977–988. [PubMed: 30103330]
14. Kramer JH, Schuff N, Reed BR, et al. Hippocampal volume and retention in Alzheimer's disease. J Int Neuropsychol Soc 2004;10(4):639–643. [PubMed: 15327742]
15. Weissberger GH, Strong JV, Stefanidis KB, Summers MJ, Bondi MW, Stricker NH. Diagnostic Accuracy of Memory Measures in Alzheimer's Dementia and Mild Cognitive Impairment: a Systematic Review and Meta-Analysis. Neuropsychology review 2017;27(4):354–388. [PubMed: 28940127]
16. Paulsen JS, Salmon DP, Monsch AU, Butters N, Swenson MR, Bondi MW. Discrimination of cortical from subcortical dementias on the basis of memory and problem-solving tests. Journal of Clinical Psychology 1995;51(1):48–58. [PubMed: 7782475]
17. Farias ST, Mungas D, Reed BR, Harvey D, Cahn-Weiner D, DeCarli C. MCI is associated with deficits in everyday functioning. Alzheimer disease and associated disorders 2006;20(4):217. [PubMed: 17132965]
18. Rog LA, Park LQ, Harvey DJ, Huang CJ, Mackin S, Farias ST. The independent contributions of cognitive impairment and neuropsychiatric symptoms to everyday function in older adults. Clin Neuropsychol 2014;28(2):215–236. [PubMed: 24502686]
19. Bäckman L, Jones S, Berger AK, Laukka EJ, Small BJ. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. Neuropsychology 2005;19(4):520–531. [PubMed: 16060827]
20. Hafiz P, Miskowiak KW, Kessing LV, et al. The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study. JMIR Form Res 2019;3(3):e13898. [PubMed: 31350840]
21. Pakhomov SV, Marino SE, Banks S, Bernick C. Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency. Speech Commun 2015;75:14–26. [PubMed: 26622073]
22. Morrison RL, Pei H, Novak G, et al. A computerized, self-administered test of verbal episodic memory in elderly patients with mild cognitive impairment and healthy participants: A randomized, crossover, validation study. Alzheimer's & dementia (Amsterdam, Netherlands) 2018;10:647–656.
23. Rey A L'examen clinique en psychologie. [The clinical examination in psychology.] Oxford, England: Presses Universitaries De France; 1964.
24. Delis DC. California verbal learning test-second edition. Adult version Manual Psychological Corporation 2000.
25. Wechsler D Wechsler memory scale-revised (WMS-R) Psychological Corporation; 1987.
26. Reitan R, Wolfson D. The Halstead-Reitan Neuropsychological Battery Theory and clinical interpretationNeuropsychology Press, Tuscan, AZ. 1993.
27. Wechsler D WAIS-IV: Wechsler adult intelligence scale San Antonio, TX: Psychological Corporation & Pearson Education, Inc.; 2008.
28. Al B, de Hamsher K. Multilingual aphasia examination Iowa City, IA. 1989.
29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin 1979;86(2):420. [PubMed: 18839484]
30. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological methods 1996;1(1):30.
31. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment 1994;6(4):284.
32. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155–163. [PubMed: 27330520]
33. Cohen J A power primer. Psychol Bull 1992;112(1):155–159. [PubMed: 19565683]
34. Schmidt M Rey auditory verbal learning test: A handbook Western Psychological Services Los Angeles, CA; 1996.
35. Heilbronner 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(8):1267–1278. [PubMed: 21108148]

36. Falleti MG, Maruff P, Collie A, Darby DG. Practice Effects Associated with the Repeated Assessment of Cognitive Function Using the CogState Battery at 10-minute, One Week and One Month Test-retest Intervals. Journal of Clinical and Experimental Neuropsychology 2006;28(7):1095–1112. [PubMed: 16840238]
37. Rast P, Macdonald SWS, Hofer SM. Intensive Measurement Designs for Research on Aging. GeroPsych (Bern) 2012;25(2):45–55. [PubMed: 24672475]
38. Sliwinski MJ. Measurement-Burst Designs for Social Health Research. Social and Personality Psychology Compass 2008;2(1):245–261.
39. Hansen TI, Lehn H, Evensmoen HR, Håberg AK. Initial assessment of reliability of a self-administered web-based neuropsychological test battery. Computers in Human Behavior 2016;63:91–97.
40. Proctor SP, Letz R, White RF. Validity of a computer-assisted neurobehavioral test battery in toxicant encephalopathy. Neurotoxicology 2000;21(5):703–714. [PubMed: 11130274]
41. Busch RM, Hogue O, Ferguson L, Parsons MW, Kubu CS, Floden DP. Validation of computerized episodic memory measures in a diverse clinical sample referred for neuropsychological assessment. The Clinical neuropsychologist 2019;33(3):557–570. [PubMed: 29996710]
42. Yonelinas AP, Widaman K, Mungas D, Reed B, Weiner MW, Chui HC. Memory in the aging brain: doubly dissociating the contribution of the hippocampus and entorhinal cortex. Hippocampus 2007;17(11):1134–1140. [PubMed: 17636547]
43. Smith PJ, Need AC, Cirulli ET, Chiba-Falek O, Attix DK. A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with “traditional” neuropsychological testing instruments. Journal of Clinical and Experimental Neuropsychology 2013;35(3):319–328. [PubMed: 23444947]
44. Tornatore JB, Hill E, Laboff JA, McGann ME. Self-administered screening for mild cognitive impairment: initial validation of a computerized test battery. The Journal of neuropsychiatry and clinical neurosciences 2005;17(1):98–105. [PubMed: 15746489]
45. Rabin LA, Paré N, Saykin AJ, et al. Differential memory test sensitivity for diagnosing amnestic mild cognitive impairment and predicting conversion to Alzheimer’s disease. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2009;16(3):357–376. [PubMed: 19353345]
46. Espinosa A, Alegret M, Valero S, et al. A longitudinal follow-up of 550 mild cognitive impairment patients: evidence for large conversion to dementia rates and detection of major risk factors involved. Journal of Alzheimer’s Disease 2013;34(3):769–780.
47. Raymond PD, Hinton-Bayre AD, Radel M, Ray MJ, Marsh NA. Test–retest norms and reliable change indices for the MicroCog Battery in a healthy community population over 50 years of age. The Clinical Neuropsychologist 2006;20(2):261–270. [PubMed: 16690546]
48. Maruff P, Thomas E, Cysique L, et al. Validity of the CogState Brief Battery: Relationship to Standardized Tests and Sensitivity to Cognitive Impairment in Mild Traumatic Brain Injury, Schizophrenia, and AIDS Dementia Complex. Archives of Clinical Neuropsychology 2009;24(2):165–178. [PubMed: 19395350]
49. Mielke MM, Machulda MM, Hagen CE, et al. Performance of the CogState computerized battery in the Mayo Clinic Study on Aging. Alzheimer’s & dementia : the journal of the Alzheimer’s Association 2015;11(11):1367–1376.
50. Weintraub S, Dikmen SS, Heaton RK, et al. Cognition assessment using the NIH Toolbox. Neurology 2013;80(11 Suppl 3):S54–S64. [PubMed: 23479546]
51. Troyer AK, Graves RE, Cullum CM. Executive functioning as a mediator of the relationship between age and episodic memory in healthy aging. Aging, Neuropsychology, and Cognition 1994;1(1):45–53.
52. Tremont G, Halpert S, Javorsky DJ, Stern RA. Differential Impact of Executive Dysfunction on Verbal List Learning and Story Recall. The Clinical Neuropsychologist 2000;14(3):295–302. [PubMed: 11262704]

53. Woods SP, Scott JC, Dawson MS, et al. Construct validity of Hopkins Verbal Learning Test—Revised component process measures in an HIV-1 sample. Archives of Clinical Neuropsychology 2005;20(8):1061–1071. [PubMed: 16198529]

54. Delis D, Kramer J, Kaplan E, Ober B. California verbal learning test–second edition. Adult version Manual The Psychological Corporation: San Antonio, TX. 2000.

55. Spreen ESEMSSO, Strauss PPE, Strauss Eet al. A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary Oxford University Press; 2006.

56. Lezak MD, Howieson DB, Bigler ED, Tranel D. Neuropsychological assessment, 5th ed. New York, NY, US: Oxford University Press; 2012.

Research in Context

Systematic Review:

The authors reviewed the literature using traditional (e.g., PubMed, PsychInfo) sources. The reliability and validity of unsupervised computerized cognitive tests for older adults is understudied, particularly for tests of learning and memory using speech recognition. However there have been several recent publications describing the psychometric properties of these types of tests. These relevant citations are provided.

Interpretation:

Our findings suggest that unsupervised computerized list learning and memory tests obtained remotely from home settings, using speech recognition is valid and has potential to significantly improve monitoring of cognitive function in older adults.

Future Directions:

Additional investigation of the reliability and validity of unsupervised cognitive tests administered remotely among older adults with neurodegenerative disease, and with respect to biomarkers of neurodegenerative disease, is an important area of research. Such studies will be important to evaluate the prognostic value using unsupervised computerized cognitive tests to identify individuals in the early stages of neurodegenerative disease.

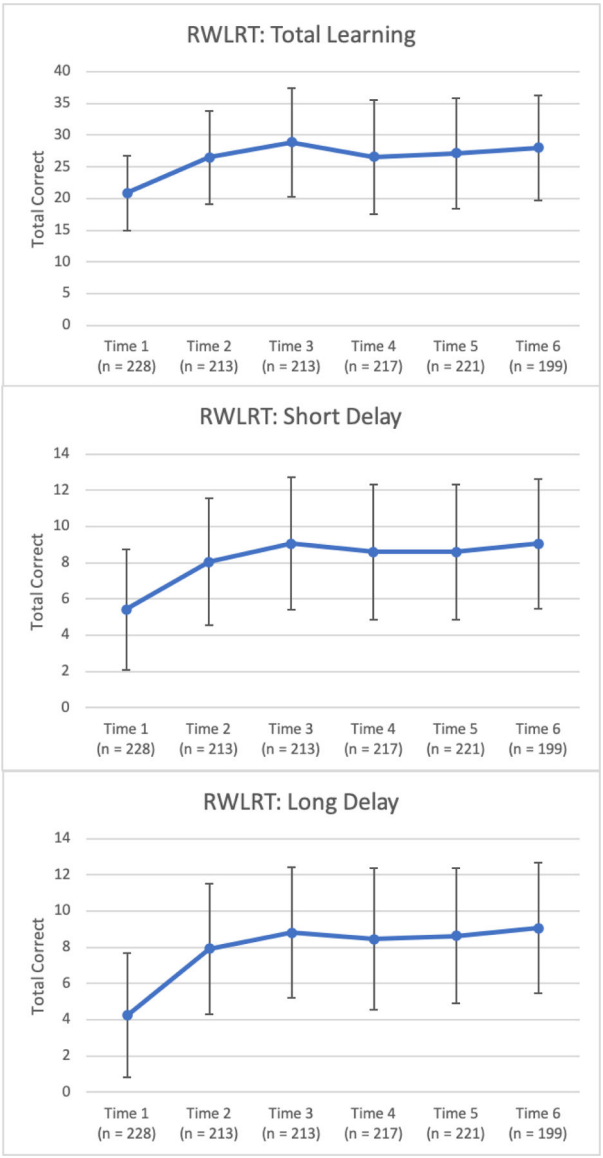

Figure 1.

Performance on RWLRT over 18 months

RWLRT = Revere Word List Recall Test

Table 1.

Sample Characteristics **(N=249)**

Variable	All participants Mean (SD)	Range
Age, mean (SD)	70.6 (6.8)	60 – 90
Years of Education, mean (SD)	16.7 (2.3)	12 – 20
Gender, n (%)		
Male	113 (51.1)	-
Female	118 (48.9)	-
Race/Ethnicity, n (%)		
Caucasian	203 (81.5)	-
Non-Caucasian	46 (18.5)	-

Table 2.

Performance on Traditional Neuropsychological Measures

Neuropsychological Test	n	Mean Raw Score	SD
CVLT-II[a] Immediate Recall	249	49.62	10.43
CVLT-II Short Delay Free Recall	249	10.39	3.18
CVLT-II Long Delay Free Recall	249	10.68	3.36
WMS-R[b] Logical Memory I	246	44.33	9.19
WMS-R Logical Memory II	246	27.87	7.55
Trail Making Test – Part A	246	33.26	11.33
Trail Making Test – Part B	245	72.59	30.11
COWAT[c]	246	45.48	12.15
WAIS-IV Digit Span	246	17.91	3.79
WAIS-IV[d] Coding	246	64.46	14.50
WAIS-IV Letter Number Sequencing	246	10.00	2.54
Grooved Pegboard – Dominant Hand	246	83.11	20.41
Grooved Pegboard – Non-dominant Hand	246	93.37	27.18

[a]CLVT-II = California Verbal Learning Test – Second Edition

[b]WMS-R = Wechsler Memory Scale – Revised

[c]COWAT = Controlled Oral Word Association Test

[d]WAIS-IV = Wechsler Adult Intelligence Scale – Fourth Edition

Table 3.

Practice Effects for RWLRT Performance Across Time Points

ReVeRe Subtest	Time 1 -Time2 Cohen's *d*	Time2 - Time3 Cohen's *d*	Time3 - Time4 Cohen's *d*	Time4 - Time5 Cohen's *d*	Time5 - Time6 Cohen's *d*
RWLRT: Total Learning	0.85	0.29	0.26	0.06	0.10
RWLRT: Short Delay	0.77	0.28	0.12	0.002	0.12
RWLRT: Long Delay	1.05	0.25	0.10	0.05	0.12

RWLRT = Revere Word List Recall Test

Table 4.

Intraclass Correlation Coefficients for Test-Retest Reliability Across Time Points

ReVeRe Subtest	All Time Points		Time 1 – 2		Time 2 – 3		Time 3 – 4		Time 4 – 5		Time 5 – 6	
	ICC (2, 6)	ICC (3, 6)	ICC (2, 1)	ICC (3, 1)	ICC (2, 1)	ICC (3, 1)	ICC (2, 1)	ICC (3, 1)	ICC (2, 1)	ICC (3, 1)	ICC (2, 1)	ICC (3, 1)
RWLRT: Total Learning	0.84	0.86	0.35	0.48	0.63	0.67	0.58	0.59	0.50	0.50	0.60	0.60
RWLRT: Short Delay	0.86	0.89	0.41	0.54	0.67	0.69	0.68	0.68	0.54	0.54	0.60	0.60
RWLRT: Long Delay	0.85	0.90	0.37	0.58	0.75	0.78	0.72	0.72	0.57	0.57	0.69	0.69

ICC = Intraclass Correlation Coefficient; RWLRT = Revere Word List Recall Test

Table 5.

Correlations of RWLRT with Traditional Neuropsychological Measures of Learning and Memory

ReVeRe Subtest	Traditional Measure	Correlation Coefficient r (*p*)
RWLRT: Total Learning	CVLT-II Immediate Recall	0.54 (<.001)
RWLRT: Short Delay	CVLT-II Short Delay Free Recall	0.59 (<.001)
RWLRT: Long Delay	CVLT-II Long Delay Free Recall	0.56 (<.001)
RWLRT: Short Delay	WMS-IV Logical Memory I	0.33 (<.001)
RWLRT: Long Delay	WMS-IV Logical Memory II	0.40 (<.001)

RWLRT = Revere Word List Recall Test; CLVT-II = California Verbal Learning Test – Second Edition; WMS-IV = Wechsler Memory Scale – Fourth Edition