Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 10 (2018) 647-656

Cognitive & Behavioral Assessment

A computerized, self-administered test of verbal episodic memory in elderly patients with mild cognitive impairment and healthy participants: A randomized, crossover, validation study

Randall L. Morrison[a,*], Huiling Pei[b], Gerald Novak[a], Daniel I. Kaufer[c], Kathleen A. Welsh-Bohmer[d], Stephen Ruhmel[e], Vaibhav A. Narayan[a]

[a]Janssen Research and Development, LLC, Titusville, NJ, USA
[b]Janssen Research and Development, LLC, Pennington, NJ, USA
[c]Department of Neurology and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
[d]Department of Neurology and Psychiatry, Duke University, Durham, NC, USA
[e]Janssen Research and Development, LLC, Raritan, NJ, USA

Abstract

Introduction: Performance of "Revere", a novel iPad-administered word-list recall (WLR) test, in quantifying deficits in verbal episodic memory, was evaluated versus examiner-administered Rey Auditory Verbal Learning Test (RAVLT) in patients with mild cognitive impairment and cognitively normal participants.

Methods: Elderly patients with clinically diagnosed mild cognitive impairment (Montreal Cognitive Assessment score 24–27) and cognitively normal (Montreal Cognitive Assessment score ≥28) were administered RAVLT or Revere in a randomized crossover design.

Results: A total of 153/161 participants (Revere/RAVLT n = 75; RAVLT/Revere n = 78) were randomized; 148 (97%) completed study; 121 patients (mean [standard deviation] age: 70.4 [7.84] years) were included for analysis. Word-list recall scores (8 trials) were comparable between Revere and RAVLT (Pearson's correlation coefficients: 0.12–0.70; least square mean difference [Revere-RAVLT]: −0.84 [90% CI, −1.15; −0.54]). Model factor estimates indicated trial ($P < .001$), period ($P < .001$) and evaluation sequence ($P = .038$) as significant factors. Learning over trials index and serial position effects were comparable.

Discussion: Participants' verbal recall performance on Revere and RAVLT were equivalent.

Keywords: Alzheimer's disease; Episodic memory; Mild cognitive impairment; Computerized assessment; Validity study

1. Introduction

Patients with mild cognitive impairment (MCI) are at increased risk of Alzheimer's disease (AD), and early therapeutic interventions at prodromal stages of AD have shown better prospects for success [1–3]. Assessment of decline in verbal episodic memory in MCI represents early cognitive changes and can be used as a screening tool for timely detection and initiation of treatment for early/preclinical AD [4]. Neuropsychological tests such as word-list recall (WLR) tests are a widely adopted approach for effective screening of cognitive abilities including memory. The Rey Auditory Verbal Learning Test (RAVLT) is a well-validated word-list based, examiner administered tool that is widely used to measure verbal episodic memory [5,6]. It provides multiple scores regarding verbal memory including rate of learning, short-term and delayed verbal memory, recall performance after interference stimulus, recognition memory, and learning pattern (serial position effect) that correlate with cognitive abilities [7,8].

Trial Registration: NCT02419183.

*Corresponding author. Tel.: +1 609 730-3620; Fax: +1 215 273-4263.
E-mail address: rmorris5@its.jnj.com

https://doi.org/10.1016/j.dadm.2018.08.010

Computerized advancements in neuropsychological assessments offer more adaptive and sensitive algorithms for detection and provide practical advantages of self-administration, automated scoring and reporting, ease of repeat adjustments, language adjustments, and reduced need for trained professionals, enabling efficient and scalable administration for large-scale screening [9–11]. The Revere software application, a computerized adaptation of RAVLT to enable self-administration and automated scoring of a WLR test, is currently being developed for use with portable electronic devices (i.e., iPad). An important first step in the use of computerized tools for cognitive assessment is their validation against the paper and pencil standard administration. The present study evaluated in-clinic feasibility and equivalence of the Revere software administration of WLR using a tablet computer to standard, examiner-administration of RAVLT using a randomized crossover design.

2. Methods

2.1. Study population

Study participants (aged 55 to 84 years) included healthy, cognitively normal (NC) individuals, and patients with MCI having normal or corrected visual and hearing acuity, as assessed clinically. Patients with clinically diagnosed MCI and Montreal Cognitive Assessment (MoCA) score (education adjusted) of 24 to 27 at study screening assessment were recruited from two specialized memory disorder clinics. Healthy NC participants with MoCA scores (education adjusted) of $\geq$28 at study screening assessment were recruited from primary care clinics [12]. Participants were English-speaking and without self-reported depressive illness as evaluated on the Patient Health Questionnaire 9-item scale (PHQ-9) [13]. Participants were evenly distributed in terms of age and sex and in the ratio of 3:1 (NC vs. MCI).

Participants were excluded if they had any acute or chronic medical, psychiatric, neurological condition or sensory, motor or speech impairment, as assessed via clinical history that would interfere with their ability to perform memory tests. Use of nicotine-based (including smoking) and caffeinated products (up to 500 mg/day) was restricted during study visits, and consumption of more than one alcoholic drink (or equivalent) within 24 hours before each WLR administration was prohibited. All concomitant medications were recorded throughout the study.

The study protocol was reviewed and approved by the Institutional Review Board at The University of North Carolina (Chapel Hill, Department of Neurology) and Duke University Medical Center. The study was conducted in accordance with the ethical principles communicated in the Declaration of Helsinki and in accordance with the International Conference on Harmonization Good Clinical Practice guidelines, applicable regulatory requirements and in compliance with the protocol. All study participants provided written informed consent before study initiation.

2.2. Study design

In this randomized, crossover, two-site study (NCT02419183) conducted in the United States from 28 May 2015 to 26 May 2016, eligible participants (both NC and patients with MCI) were allocated (1:1) to receive either Revere or RAVLT in period 1 and then crossed-over to the alternate test in period 2, with a memory washout interval of 7 to 14 days. In doing so, participants assignments were rigorously counterbalanced with 50% of the study population receiving the computerized administration of the RAVLT first and the other 50% receiving the standard in person assessment with the RAVLT first. An optional follow-up visit was scheduled within 7 days after the last study-related activity for participants who experienced an adverse event (AE) during the study that was unresolved by the end of the last test visit (Fig. 1).

The RAVLT was administered with five presentations and recall attempts of a 15-word list (Word-list A), a distractor task (Word-list B), post-distractor recall (Word-list A), and a 20-minute delayed recall (Word-list A). All trials involved the examiner reading the word-list aloud to the participants except the post-distractor and 20-minute delayed recall assessments. Score for each recall trial was the number of words correctly recalled by the participant. The Revere software administration of WLR mimicked the standard RAVLT administration; instructions for the Revere WLR test were provided to the participant via iPad, both aloud and on screen. Word-lists were "read" aloud by the software, in a manner mimicking standard administration of the RAVLT; that is, with five presentations and recall attempts of a 15-word-list (Word-list A), a distractor task (Word-list B), post-distractor recall of Word-list A, and a 20-minute delayed recall of Word-list A. After presentation of each list by the software, the participant was asked to recall as many words from the list that he/she had just heard read aloud. The post-distractor and 20-minute delayed recall assessments were conducted without the software reading Word-list A aloud.

The Revere software recorded participants' recall responses; performance on each WLR trial was subsequently scored by human rater(s) and automated speech recognition software (details to be elaborated elsewhere) for number of words correctly recalled for each recall trial. Revere WLR and examiner-administered RAVLT were run under the same conditions. Assessments were conducted in a quiet clinic room. Before administering the WLR test, the Revere software asked participants to read a paragraph aloud and to complete a brief digit-span task (recalling a 4-digit number as presented aloud by the software followed by a 5- and a 6-digit number). These tasks were administered as control

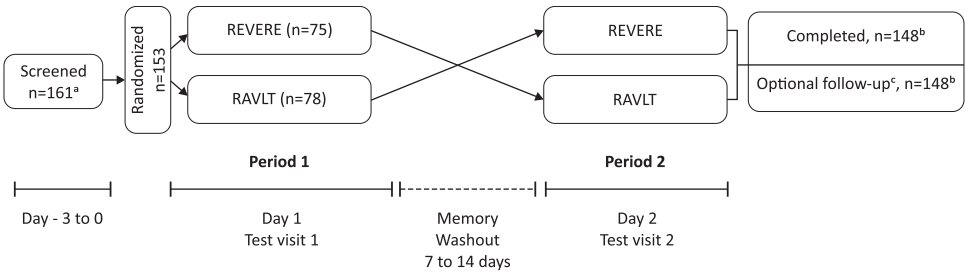


Fig. 1. Study design and patient flow; [a]n = 8 participants excluded (screen failures); [b]n = 5 participants withdrew consent; [c]optional follow-up visit within 7 days of the last test visit was scheduled for participants who experienced an adverse event that had not resolved till the test visit 2. Abbreviations: RAVLT, Rey Auditory Verbal Learning Test (conventional examiner-administered word-list recall test); Revere, Self-Administered Memory Screening Test with Automated Reporting (iPad computer-administered word-list recall test).

tasks to acclimate the participant to the test conditions using the iPad. Participants progressed on the digit span to the next digit recall if they succeeded at prior digit length.

During the delay-period before the 20-minute delayed recall assessment, participants were provided with travel magazine articles and asked to read through three articles of interest, at the end of which they were asked by a study research assistant to briefly describe two of the articles (i.e., travel destination that was the focus of the article and two landmarks/attractions per destination); their responses were transcribed. Twenty minutes after completion of the list recall tasks, the participants were asked to recall as many words as possible from the original word-list (Word-list A).

Recall on learning trials of the original word-list, recall of the original word-list following the distractor task, and delayed recall were the primary outcome measures. Audio recordings of the participant responses to the Revere WLR test were scored by two independent raters. The final consensus score from the two raters was used for the primary analysis of the study.

2.3. Participant survey and patient-reported outcome

All participants and attending clinicians or their staff completed a survey after WLR tests in period 2 regarding their attitudes about cognitive screening and receptivity to a computerized cognitive assessment and acceptance after familiarization with the computerized test versus examiner-based cognitive tests. The clinic staff and physicians were also asked to provide more detailed feedback about potential advantages and disadvantages of different screening methods to assess the feasibility of a self-administered cognitive assessment software via tablet computer in primary care and specialty practices. Participants also completed the PHQ-9 at screening and after the WLR tests in both periods 1 and 2.

2.4. End points

2.4.1. Primary end point

Primary end point was the number of words successfully recalled from the Revere software-administered WLR and RAVLT on trials I to V (Word-list A), distractor test (Word-list B), post-distractor recall (Word-list A), and the 20-minute delayed recall (Word-list A).

2.4.2. Secondary end points

The secondary end points included further analysis of WLR test scores to assess total learning (total score from trials I to V), learning over trials index ([total learning from trial I to V]$-5\times$ [score from trial I]) and serial position analysis of recalled words (primacy [percentage of words, positioned 1 through 5 on the list, successfully recalled across trials I to V], middle region [percentage of words, positioned 6 through 10 on the list, successfully recalled across trials I to V], and recency [percentage of words, positioned 11 through 15 on the list, successfully recalled across trials I to V]).

2.4.3. Safety

Safety assessments included monitoring of AEs, vital signs, and physical examination.

2.5. Statistical methods

2.5.1. Sample size

Considering a within-participant correlation of 0.61, the sample size of 96 evaluable participants provided 80% power to detect equivalence using a single-word (7%) equivalence margin on any WLR trial between the computerized Revere WLR test and examiner-administered RAVLT. Participants with data losses due to computer or software-related glitches were replaced to ensure that data from at least 96 participants are available for analysis.

2.5.2. Analysis sets

The evaluable analysis set included all participants who were randomized and had WLR scores for Revere (consensus scores from two independent rates) and RAVLT tests. Randomized participants analysis set included all participants randomized to the WLR tests.

2.5.3. Assessments and primary analysis

The performances on all trials based on WLR scores (primary end point) were compared between Revere and RAVLT

using Schuirmann's two one-sided test for equivalence. Primary analysis to establish equivalence between Revere (based on independent rater scorings of audio recordings) and RAVLT was performed using repeated measure mixed model, with WLR scores from the eight trials as dependent variable; period, evaluation sequence, evaluation group (computer or examiner), trials (I to VIII), interaction of evaluation group and trial as fixed effects and participants as random effect.

One-half of standard deviation (SD) is regarded as a widely accepted approach to interpret clinically meaningful differences between cognitive abilities measured using neuropsychological tests [14]. According to the published norm of RAVLT, SDs from various trials in elderly participants ranged from 1.9 to 3.4 with the median SD of 2.7 [5]. Therefore, to determine the magnitude of difference in cognitive performance between Revere and RAVLT, the clinically relevant equivalence margin was set to 1.35 (one-half of 2.7). Equivalence was established if the 90% confidence intervals (CIs) of least square mean difference were well within the prespecified range of (−1.35, 1.35). The WLR scores were also presented descriptively (by evaluation group and trial).

WLR scores were summarized based on the NC and MCI subgroups, gender, subgroups (men, women), and four age subgroups for NC participants as presented in the RAVLT manual (50–59 years, 57 to 69 years, 70 to 79 years, and 76 to 89 years) [5].

2.6. Sensitivity analyses for primary end point

Sensitivity analyses were performed to further test equivalence between the computerized Revere and examiner-administered RAVLT: (1) using the same primary analysis model with only scores from trials I to V as dependent variable; (2) using similar mixed model as the primary analysis by adding key demographic and stratification factors (NC/MCI status, age [55 to 64 years, 65 to 74 years, 75 to 84 years] and gender) as fixed effects; and (3) repeated measure analysis using WLR scores from automated scoring (speech recognition engine [SRE]: Nuance Speech Anywhere) for the Revere WLR test as dependent variable.

Performances on the Revere computerized WLR test and RAVLT were also compared based on the total scores from trials I to V and the 20-minute delayed recall test using Bland-Altman plots and Deming regression.

3. Results

3.1. Demographics and baseline characteristics

A total of 148 of 153 (97%) randomized participants completed the entire crossover study (Fig. 1). Five participants discontinued from the study (three in Revere/RAVLT sequence [2 NC, 1 MCI] and two in RAVLT/Revere sequence [one NC, one MCI], all five: withdrawal of consent). Of the 153 randomized participants, 32 were excluded due to unavailability of audio recordings from the Revere (irretrievable audio recordings, n = 25; withdrawal of consent, n = 2; iPad malfunctions, n = 5). Of note, a total of 17 (11.1%) participants had iPad issues either due to iPad malfunctions or user error, of which, five resulted in loss of computer assessment data and were replaced; the remaining participants had one or two missing trial data from the computer assessment. The evaluable analysis set included 121 randomized participants.

The mean (SD) age of participants was 70.4 (7.84) years, and the majority were women (84/121 [69.4%]). Most participants had normal cognition at study entry (94/121 [77.7%] had MoCA ≥28 and clinical diagnosed as NC). At baseline, based on the PHQ-9 total score, most of the participants (103/121 [85.1%]) had minimal or no symptoms of depression, while 15/121 (12.4%) had mild, 2/121 (1.7%) had moderate and 1/121 (0.8%) had moderately severe symptoms of depression (Table 1). Participants provided information on their educational level in the participant study survey (n = 134): master's degree, 61 (45.5%); bachelor's degree, 34 (25.4%); doctorate degree, 20 (14.9%); professional school degree, 7 (5.2%); associate's degree, 5 (3.7%); high school, 4 (3.0%); and some college, 3 (2.2%).

Most of the randomized participants (92%) were taking medications before and during study participation. The most common prior and concomitant medications taken by at least 30% of participants included vitamin D and analogues, platelet aggregation inhibitors (excluding heparin), and multivitamins. Less than 15% of participants took antidepressant medications, and 5% or fewer of participants took benzodiazepine-related drugs (sleep aids) or benzodiazepine derivatives before and during the study. No participant reported consumption of more than one alcoholic drink within 24 hours before WLR administration or exceeded the daily limit of 500 mg caffeine since the study screening visit.

3.2. Primary end point

The mean (SD) WLR scores increased from trials I to V in both computer-administered Revere (6.0 [2.03] to 11.2 [2.51]) and examiner-administered RAVLT tests (6.4 [2.13] to 12.1 [2.35]), with the examiner-administered RAVLT showing slightly better performances across the eight trials than Revere (Fig. 2). Pearson's correlation coefficients between Revere and RAVLT for the eight trials ranged from 0.12 to 0.70, with higher level of correlation for the later trials on word-list A.

From the repeated measure mixed model for scores obtained from the eight trials, the least square mean difference between Revere and RAVLT was −0.84 (90% CI: −1.15, −0.54), which was fully contained in the prespecified equivalence range of (−1.35, 1.35) with significant equivalence (equivalence margin adjusted P value = .003) (Table 2). The overall equivalence test between Revere and RAVLT was the confirmatory test, and results from individual trials were supportive. Therefore, multiplicity adjustment was not needed. Significant factors in repeated measure mixed

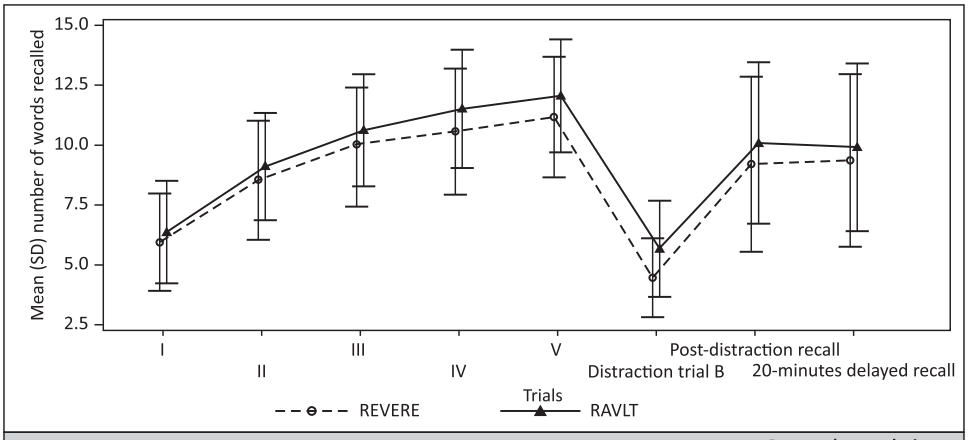
Table 1
Demographics and baseline characteristics (evaluable analysis set)

Characteristics	Revere/RAVLT (n = 56)	RAVLT/Revere (n = 65)	Total (n = 121)
Disease status, n (%)			
Mild cognitive impairment*	13 (23.2)	14 (21.5)	27 (22.3)
Normal control	43 (76.8)	51 (78.5)	94 (77.7)
Age, mean (SD), years	70.1 (7.34)	70.7 (8.28)	70.4 (7.84)
Gender, n (%)			
Women	39 (69.6)	45 (69.2)	84 (69.4)
Race, n (%)			
White	51 (91.1)	61 (93.8)	112 (92.6)
Black or African-American	5 (8.9)	2 (3.1)	7 (5.8)
Asian	0	2 (3.1)	2 (1.7)
PHQ-9 total score, n (%)			
Severe depression (>20)	0	0	0
Moderately severe depression (15–19)	0	1 (1.5)	1 (0.8)
Moderate depression (10–14)	1 (1.8)	1 (1.5)	2 (1.7)
Mild depression (5-9)	9 (16.1)	6 (9.2)	15 (12.4)
No or minimal depression (0–4)	46 (82.1)	57 (87.7)	103 (85.1)
Mean (SD)	2.2 (2.52)	2.3 (3.09)	2.2 (2.83)
MoCA total score, n (%)			
Severe cognitive impairment (<10)	0	0	0
Moderate cognitive impairment (10–17)	0	0	0
Mild cognitive impairment (18–26)	9 (16.1)	8 (12.3)	17 (14.0)
No cognitive impairment (>26)	47 (83.9)	57 (87.7)	104 (86.0)
Mean (SD)	28.2 (1.67)	28.2 (1.39)	28.2 (1.52)

Abbreviations: MoCA, Montreal Cognitive Assessment; PHQ-9, Patient Health Questionnaire (9 item); RAVLT, Rey Auditory Verbal Learning Test; Revere, Self-Administered Memory Screening Test with Automated Reporting; SD, standard deviation.
*Clinically diagnosed and based on MoCA total scores.

model testing were trial (suggesting learning effect from trial to trial), period ($P < .001$ for both), and evaluation sequence ($P = .038$) (detailed in Supplementary Table A1 of Supplementary Appendix A).

Sensitivity analysis using the primary analysis model for WLR scores from the learning trials (trials I to V) and inclusion of demographic stratification supported the equivalence claim for the Revere and RAVLT tests. Automated WLR



Trial	REVERE (n=121)	RAVLT (n=121)	Pearson's correlation (coefficient [95% CI])
Trial I	6.0 (2.03)[a]	6.4 (2.13)	0.12 (−0.06, 0.30)
Trial II	8.5 (2.48)	9.1 (2.25)	0.45 (0.29, 0.58)
Trial III	10.0 (2.35)[a]	10.6 (2.34)	0.62 (0.50, 0.72)
Trial IV	10.6 (2.64)	11.5 (2.47)	0.61 (0.49, 0.71)
Tria IV	11.2 (2.51)	12.1 (2.35)	0.67 (0.56, 0.76)
Distraction trial (word list B)	4.5 (1.66)[b]	5.7 (2.01)	0.40 (0.24, 0.54)
Post-distraction immediate recall	9.2 (3.66)	10.1 (3.36)	0.70 (0.60, 0.78)
20-minute delayed recall	9.4 (3.60)[c]	9.9 (3.50)	0.63 (0.51, 0.73)

Fig. 2. Mean word-list recall test scores (Evaluable analysis set); Abbreviations: CI, confidence interval; RAVLT, Rey Auditory Verbal Learning Test; Revere, Self-Administered Memory Screening Test with Automated Reporting; NOTE. All values are expressed as mean (SD). [a]n = 120; [b]n = 117; [c]n = 119.

Table 2
Primary analysis: Repeated measure mixed model of world list recall test scores (evaluable analysis set)

	Revere (n = 121)	RAVLT (n = 121)	Difference (Revere − RAVLT)		
Trial	LS mean (SE)	LS mean (SE)	LS mean (SE)	90% CI	*P* value*
Trial I	5.92 (0.22)	6.40 (0.22)	−0.49 (0.24)	(−0.88, −0.09)	<.001
Trial II	8.48 (0.22)	9.14 (0.22)	−0.66 (0.24)	(−1.05, −0.26)	.002
Trial III	9.98 (0.22)	10.64 (0.22)	−0.66 (0.24)	(−1.05, −0.26)	.002
Trial IV	10.51 (0.22)	11.53 (0.22)	−1.03 (0.24)	(−1.42, −0.63)	.089
Trial V	11.11 (0.22)	12.08 (0.22)	−0.97 (0.24)	(−1.36, −0.58)	.056
Distraction trial (word-list B)	4.39 (0.22)	5.69 (0.22)	−1.30 (0.24)	(−1.70, −0.90)	.417
Post-distraction immediate recall	9.14 (0.220)	10.11 (0.22)	−0.97 (0.24)	(−1.36, −0.58)	.056
20-minute delayed recall	9.26 (0.22)	9.94 (0.22)	−0.67 (0.24)	(−1.07, −0.28)	.003
Overall	8.60 (0.17)	9.44 (0.17)	−0.84 (0.18)	(−1.15, −0.54)	.003

Abbreviations: CI, confidence interval; LS, least square; RAVLT, Rey Auditory Verbal Learning Test; Revere, Self-Administered Memory Screening Test with Automated Reporting; SE, standard error.

NOTE. Analysis was based on a repeated measure mixed model for the number of words successfully recalled from all trials (acquisition trials I to V, distraction Trial B, post-distraction recall and 20 minute delayed recall) as dependent variable, and period, evaluation sequence, evaluation group (Revere or RAVLT), trial, interaction of evaluation group and trial as fixed effects and participants as random effect.

*Equivalence margin adjusted *P* value.

scoring using the Nuance Speech Anywhere SRE was accurate (>97%). Participant's performance on Revere software as scored by the SRE was equivalent to performance on examiner-administered RAVLT ($P = .050$). The overall least square mean difference between Revere with SRE scoring and RAVLT was −1.03 (90% CI: −1.35, –0.72), falling within the prespecified range of (−1.35, 1.35) (Table 3). The Deming regression for WLR scores from trials I to V showed slope of 1.13 (95% CI: 0.91, 1.36) and intercept of −10.21 (95% CI: −21.54, 1.12) and supported the agreement in performances of Revere and RAVLT (Supplementary Fig. B1, Supplementary Appendix B).

3.3. Secondary end points

The mean (SD) of learning over trial indices were comparable between Revere and RAVLT (16.4 [6.77] vs. 17.8 [8.22]; Pearson's correlation coefficient [95% CI]: 0.38 [0.22, 0.53]). Similar results were observed for the serial position analysis. Primacy and recency effects were seen for both Revere and RAVLT. Similar percent of words were recalled from the initial part of the list (primacy region) for both Revere and RAVLT, while the percent of words recalled from the middle (middle [% recalled] mean [SD]: 50.9 [18.86] vs. 56.4 [19.57]) and end of the list (recency region) was greater for RAVLT than Revere (recency [% recalled] mean [SD]: 65.7 [15.63] vs. 73.5 [14.12]). Pearson's correlation coefficients between Revere and RAVLT for serial position effects ranged from 0.37 to 0.56, with primacy region having the highest correlation (Table 4).

3.4. Subgroup analyses

In the subgroup analysis comparing mean WLR scores from eight trials, as anticipated, the mean number of words recalled was higher for NC participants versus patients with MCI. With regard to serial position effect, patients with MCI had lower mean WLR scores suggesting reduced primacy and recency effects as compared with NC participants. Similarly, in the age-modified WLR performance the NC

Table 3
Sensitivity analysis (evaluable analysis set)

	Revere (n = 121)	RAVLT (n = 121)	Difference (Revere − RAVLT)		
Trial	LS mean (SE)	LS mean (SE)	LS mean (SE)	90% CI	*P* value*
Repeated measure of WLR scores from trials I to V					
Overall	9.20 (0.16)	9.96 (0.17)	−0.76 (0.19)	(−1.07, −0.45)	.001
Repeated measure of WLR scores from all trials with demographic factors					
Overall	8.01 (0.19)	8.85 (0.19)	−0.84 (0.17)	(−1.13, −0.56)	.002
Repeated measure of WLR scores from all trials based on Nuance Speech Recognition System					
Overall	8.41 (0.17)	9.44 (0.17)	−1.03 (0.19)	(−1.35, −0.72)	.050

Abbreviations: CI, confidence interval; LS, least square; RAVLT, Rey Auditory Verbal Learning Test; Revere, self-administered memory screening test with automated reporting; SE, standard error; WLR, word-list recall.

NOTE. Analysis was based on a repeated measure mixed model for the number of words successfully recalled from the trials (acquisition trials I to V, distraction Trial B, post-distraction recall and 20-minute delayed recall) as dependent variable, and period, evaluation sequence, evaluation group (Revere or RAVLT), trial, interaction of evaluation group and trial as fixed effects and participant as random effect.

*Equivalence margin adjusted *P* value.

Table 4
Word-list recall test scores for secondary end points (evaluable analysis set)

Index	Revere (n = 121) mean (SD)	RAVLT (n = 121) mean (SD)	Pearson's correlation coefficient (95% CI)
Total (trials I to V)	46.1 (10.61)*	49.7 (9.75)	0.62 (0.50, 0.72)
LOT index†	16.4 (6.77)*	17.8 (8.22)	0.38 (0.22, 0.53)
Serial position effect (Trials I to V)‡			
Primacy (% recalled)	68.8 (18.25)	69.0 (16.31)	0.56 (0.43, 0.67)
Middle region (% recalled)	50.9 (18.86)	56.4 (19.57)	0.48 (0.33, 0.61)
Recency (% recalled)	65.7 (15.63)	73.5 (14.12)	0.37 (0.20; 0.51)

Abbreviations: CI, confidence interval; LOT, learning over trials; RAVLT, Rey Auditory Verbal Learning Test; Revere, Self-Administered Memory Screening Test with Automated Reporting.

*n = 119.

†LOT index = Total of trials (I to V)–(5X [trial I]).

‡Percent recalled from each region (primacy, middle region, recency) = number of words successfully recalled from the region divided by the total number of words presented in that region from the word-list with the five learning trials (trials I to V) combined. (primacy region, words from positions 1 to 5 of the word-list; middle region, words from positions 6 to 10; recency region, words from positions 11 to 15).

participants from the youngest age subgroup (50–59 years) had higher mean WLR scores and out-performed participants from the older age groups. However, a gender effect was notably absent, and no performance differences in WLR scores were observed between men and women.

Primary and recency effects were noted in all age subgroups. In subgroup analysis of gender difference, mean words recalled by men and women were consistent with the overall population. Performance with RAVLT was slightly better than Revere, irrespective of gender. The Pearson's correlation analysis in both men (Pearson's correlation coefficient, 0.24 [trial I] to 0.60 [20-minute delayed recall]) and women (0.07 [trial I] to 0.73 [post distraction immediate recall]) suggested high levels of correlation between the two tests. Subgroup analysis has been detailed in Supplementary Appendix C (Supplementary Tables C1 to C8).

PHQ-9 scores and participant survey report are summarized in Supplementary Appendix D.

3.5. Safety

A total of, 7 of 153 (5%) participants experienced AEs (MCI: sciatic pain, concussion without loss of consciousness, chest infection; NC: back spasm, worsening back pain, viral infection, and cutaneous small B lymphoma, n = 1 each). All AEs except the small B cell lymphoma were mild or moderate in severity. All observed AEs appeared to be unrelated to study participation and were common occurrences for this population. No AE led to study discontinuation. No clinically relevant changes from baseline were noted in vital sign measurements.

4. Discussion

Development of computerized cognitive tests is of huge importance in perpetuating clinical trials productively in the preclinical AD space [15,16]. The approach helps in improving the scientific quality of the data by eliminating common sources of examiner errors in administration and scoring, thereby allowing greater sensitivity and more fidelity for the detection and monitoring of clinical change and disease progression. These computerized methods allow for controlled timing between trials and permit data capture of a host of measures, such as response times, which are not easily captured with standard administration. However, with the introduction of an additional digital interface, it is important to minimize possible interference with user's cognitive faculties that can potentially influence results. Thus, there is an increasing emphasis on validation and assessment of psychometric properties of computerized tests to minimize the risk of conflicting results and maintain equivalence between traditional and digital formats [11]. From a clinical and research standpoint, validated computerized tests along with their equivalent examiner-based tests offer practicality and flexibility to use either format that may increase the number of individuals being screened.

The present study was conducted to evaluate the psychometric properties (criterion validity) of Revere, a computerized adaptation of the RAVLT in NC participants and patients with MCI, against a standard version of the RAVLT, administered by an examiner under the same in-clinic conditions. Participant performance on the Revere demonstrated equivalence to participant performance on the RAVLT based on scores from eight trials, using an equivalence margin determined as 0.5 SD based on age-adjusted normative data. Sensitivity analysis conducted for the WLR scores using only scores from learning trials I to V or a repeated measures mixed model adding the stratification factors of the participants as fixed effects in the model further corroborated these findings. Automated scoring by the Nuance Speech Anywhere SRE demonstrated high level of accuracy and participant performance based on automated scoring on Revere was equivalent to the performance on RAVLT, suggesting prospects for fully-computerized cognitive assessments.

Consistent with published literature for examiner-administered WLR tests, a learning effect over trials was

observed in this study [7]. The test-retest improvement is anticipated in healthy adult populations on verbal list learning tests, presumably reflecting increased comfort and familiarity with the memory testing procedures [17]. The practice effect was seen regardless of the order of test administration (i.e., iPad administration first vs. standard administration first). Mean WLR scores from trials I to V increased for both Revere as well as RAVLT with same pattern demonstrated across the eight trials in both the groups although, RAVLT showed slightly better performance than Revere. When the factors included in the repeated measure mixed model were tested, the "trial" factor was statistically significant; this indicates a learning effect from trial to trial. The mean scores for each individual trial for both Revere and RAVLT from period 2 were higher than the corresponding mean scores from period 1 tests, thus indicating a learning effect from period 1 to period 2 despite the 7- to 14-day memory washout between the two periods.

Serial position effects are commonly observed in free recall, and decreased recall of primacy words has been associated with prediction of cognitive decline [18,19]. Consistent with this, participant performance on Revere and RAVLT showed more words correctly recalled from the primacy and recency positions of the word-list than the middle region. Although trends in patients with MCI were similar to the overall population, patients with MCI recalled fewer words than the NC subgroup and had reduced primacy and recency effects. There was no meaningful impact of gender on cognitive performance measure with the use of either Revere or RAVLT, consistent with previous observations of neurocognitive measures [20]. Age had negligible influence and younger participants performed better across trials and had superior primacy and recency effects. No safety concerns were identified with the use of Revere.

A significant challenge with the introduction of newer technology such as computer- or touch screen device-based testing modality is the variability in the level of preference or acceptability among older adults due to lack of experience, low familiarity, or general perception [21,22]. In the present study, elderly patients showed favorable acceptability and were comfortable with iPad-based interaction, and most participants were willing to take future Revere tests. This observation adds to available data that support the use of automated, computerized cognitive batteries in elderly individuals [21–25]. Data loss from failure to capture or submit scorable audio recordings due to technical glitches in the software or network issues was another notable challenge. The Revere software was updated during the trial to prevent application crashes and the sites were moved to more robust wireless networks to prevent data transmission errors. Data loss due to human error was addressed by ensuring consistent backup of all data files and later altered to use a new, more robust data storage mechanism.

The study strengths included methodologic rigor (crossover design), relatively large sample size, and recruitment of elderly patients from clinical care sites including a Memory Disorders Clinic. Further, the approach to cognitive screening of MCI was based on a combination of clinical diagnosis and a MoCA cut-off $\geq$28 to allow inclusion of early-stage MCI. A limitation of the study is that the memory washout between the test periods was not long enough to rule out learning/practice effects contributing to better performance observed during period 2. The overall level of education in this sample was high (>90% with college education or greater). Performance in less well-educated cohorts is not yet completely known, however, we note that the range of values was good across the MCI and older control cohorts without evidence of ceiling or floor effects in either type of administration. Finally, participants with MCI were diagnosed based on clinical assessment for their memory complaints; there was no biomarker confirmation of preclinical AD.

In conclusion, the present study demonstrated equivalence between the iPad-administered Revere WLR test and examiner-administered RAVLT in terms of the participant's verbal recall performance. Revere provides a comprehensive profile of cognitive abilities related to verbal episodic memory, including learning and delayed recall. Thus, computerized measurements of cognitive attributes using tools such as Revere may enhance the prospects of yielding standardized, sensitive and reproducible end points, assisting large-scale screening during clinical research and optimizing treatment choices in mainstream clinical practice. Based on the performance equivalence of Revere and RAVLT, further development and validation of a completely automated version of Revere with speech recognition technology that is functional in a home-based setting is underway. Home-based implementation of a validated automated cognitive tool would potentially facilitate unsupervised and reliable collection of longitudinal information on memory and cognitive abilities of elderly individuals within a real-world environment. The current version of Revere can be made available on request to support joint data collection that would contribute to further validation of the system in real-world settings.

Acknowledgments

The authors thank Shruti Shah, PhD and Priya Ganpathy, MPharm, ISMPP CMPP^{TM} (SIRO Clinpharm Pvt. Ltd.) for writing assistance and Ellen Baum, PhD (Janssen Research & Development, LLC) for additional editorial assistance. The authors also thank the study participants, without whom this study would not have been accomplished, as well as the investigators for their participation in this study.

Funding: The study is sponsored by Janssen Research & Development, LLC, USA.

Author contributions: R.L.M., G.N., S.R., and V.A.N. were involved in study design, data collection, analysis, and interpretation. H.P. was responsible for statistical analyses and

design/interpretation of study results. D.I.K. and K.A.W.-B. were principal investigators for this study. All authors had full access to study data and take responsibility for the integrity of data and accuracy of data analysis. All authors meet ICMJE criteria and all those who fulfilled those criteria are listed as authors. All authors provided direction and comments on the manuscript, made the final decision about where to publish these data, and approved submission to this journal.

Declaration of Interest: R.L.M., H.P., G.N., S.R., and V.A.N. are full time employees of Janssen Research & Development, LLC. D.I.K. and K.A.W.B. were principal investigators for this study and have received research support or fees as consultants for Janssen Research & Development, LLC.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dadm.2018.08.010.

RESEARCH IN CONTEXT

1. Systematic review: We reviewed the literature using PubMed for relevant reports on screening tools for mild cognitive impairment. Impaired verbal episodic memory is regarded as a predictive marker of mild cognitive impairment and its reliable detection is crucial for diagnosing prodromal phase Alzheimer's disease.
2. Interpretations: Recall performance using the iPad-based Revere (a computerized adaptation of the Rey Auditory Verbal Learning Test) was equivalent to the conventional examiner-based Rey Auditory Verbal Learning Test module. Revere enabled comprehensive assessment of verbal episodic memory comprising immediate and delayed recall, learning effect and serial position effect. Automated scoring with Revere demonstrated high accuracy and equivalence to Rey Auditory Verbal Learning Test scoring.
3. Future directions: Revere improves prospects for developing more adaptive, sensitive, reproducible and standardized screening algorithm for large-scale screening of cognitive defects. Further development of a completely automated version of Revere is underway that would potentially enable unsupervised administration in home-based setting and assist reliable collection of longitudinal data.

References

[1] Meyer JS, Xu G, Thornby J, Chowdhury MH, Quach M. Is mild cognitive impairment prodromal for vascular dementia like Alzheimer's disease? Stroke 2002;33:1981–5.

[2] Petersen RC, Stevens JC, Ganguli M, Tangalos EG, Cummings JL, DeKosky ST. Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2001;56:1133–42.

[3] Budd D, Burns LC, Guo Z, L'Italien G, Lapuerta P. Impact of early intervention and disease modification in patients with predementia Alzheimer's disease: A Markov model simulation. Clinicoecon Outcomes Res 2011;3:189–95.

[4] Leube DT, Weis S, Freymann K, Erb M, Jessen F, Heun R, et al. Neural correlates of verbal episodic memory in patients with MCI and Alzheimer's disease–a VBM study. Int J Geriatr Psychiatry 2008; 23:1114–8.

[5] Schmidt M. Rey Auditory and Verbal Learning Test: A handbook. Los Angeles, CA: Western Psychological Services; 1996.

[6] Schoenberg MR, Dawson KA, Duff K, Patton D, Scott JG, Adams RL. Test performance and classification statistics for the Rey Auditory Verbal Learning Test in selected clinical samples. Arch Clin Neuropsychol 2006;21:693–703.

[7] Estevez-Gonzalez A, Kulisevsky J, Boltes A, Otermin P, Garcia-Sanchez C. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: Comparison with mild cognitive impairment and normal aging. Int J Geriatr Psychiatry 2003;18:1021–8.

[8] Ranjith N, Mathuranath PS, Sharma G, Alexander A. Qualitative aspects of learning, recall, and recognition in dementia. Ann Indian Acad Neurol 2010;13:117–22.

[9] Snyder PJ, Jackson CE, Petersen RC, Khachaturian AS, Kaye J, Albert MS, et al. Assessment of cognition in mild cognitive impairment: a comparative study. Alzheimers Dement 2011; 7:338–55.

[10] Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. Clin Neuropsychol 2012;26:177–96.

[11] Ruggeri K, Maguire A, Andrews JL, Martin E, Menon S. Are We There Yet? Exploring the impact of translating cognitive tests for dementia using mobile technology in an aging population. Front Aging Neurosci 2016;8:21.

[12] Nasreddine ZS, Phillips NA, Bedirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005;53:695–9.

[13] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–13.

[14] Harvey PD. Clinical applications of neuropsychological assessment. Dialogues Clin Neurosci 2012;14:91–9.

[15] Gualtieri CT, Johnson LG. Neurocognitive testing supports a broader concept of mild cognitive impairment. Am J Alzheimers Dis Other Demen 2005;20:359–66.

[16] Daffner KR, Gale SA, Barrett AM, Boeve BF, Chatterjee A, Coslett HB, et al. Improving clinical cognitive testing: report of the AAN Behavioral Neurology Section Workgroup. Neurology 2015; 85:910–8.

[17] Benedict R, Schretlen D, Groninger L, Brandt J. Hopkins Verbal Learning Test – Revised: Normative Data and Analysis of Inter-Form and Test-Retest Reliability. Clin Neuropsychol 1998;12:43–55.

[18] Bruno D, Reiss PT, Petkova E, Sidtis JJ, Pomara N. Decreased recall of primacy words predicts cognitive decline. Arch Clin Neuropsychol 2013;28:95–103.

[19] La Rue A, Hermann B, Jones JE, Johnson S, Asthana S, Sager MA. Effect of parental family history of Alzheimer's disease on serial position profiles. Alzheimers Dement 2008;4:285–90.

[20] Welsh-Bohmer KA, Ostbye T, Sanders L, Pieper CF, Hayden KM, Tschanz JT, et al. Neuropsychological performance in advanced age: influences of demographic factors and Apolipoprotein E: Findings from the Cache County Memory Study. Clin Neuropsychol 2009; 23:77–99.

[21] Canini M, Battista P, Della Rosa PA, Catricala E, Salvatore C, Gilardi MC, et al. Computerized neuropsychological assessment in aging: Testing efficacy and clinical ecology of different interfaces. Comput Math Methods Med 2014;2014:804723.

[22] Wood E, Willoughby T, Alice R, Bechtel L, Gilbert J. Use of Computer Input Devices by Older Adults. J Appl Gerontol 2005;24:419–38.

[23] Kaye J, Mattek N, Dodge HH, Campbell I, Hayes T, Austin D, et al. Unobtrusive measurement of daily computer use to detect mild cognitive impairment. Alzheimers Dement 2014;10:10–7.

[24] Rentz DM, Dekhtyar M, Sherman J, Burnham S, Blacker D, Aghjayan SL, et al. The Feasibility of At-Home iPad Cognitive Testing For Use in Clinical Trials. J Prev Alzheimers Dis 2016;3:8–12.

[25] Sano M, Egelko S, Ferris S, Kaye J, Hayes TL, Mundt JC, et al. Pilot study to show the feasibility of a multicenter trial of home-based assessment of people over 75 years old. Alzheimer Dis Assoc Disord 2010;24:256–63.