Title:

Validation of a novel fully automated story recall task for repeated remote high-frequency administration

Author names:
Caroline Skirrow*[1], Marton Meszaros[1], Udeepa Meepegama[1], Raphael Lenain[1], Kathryn V. Papp[2,3], Jack Weston[1], Emil Fristed[1]

Affiliations:
[1]Novoic Ltd, London, UK.
[2]Center for Alzheimer Research and Treatment, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts
[3]Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston

*corresponding author: caroline@novoic.com

Abstract:

INTRODUCTION: Longitudinal data is key to identifying cognitive decline and treatment response in Alzheimer's disease (AD).

METHODS: The Automatic Story Recall Task (ASRT) is a novel, fully automated test that can be self-administered remotely. In this longitudinal case-control observational study, 151 participants (mean age: 69.99 (range 54-82), 73 mild cognitive impairment/mild AD and 78 cognitively unimpaired) completed parallel ASRT assessments on their smart devices over 7-8 days. Responses were automatically transcribed and scored using text similarity metrics.

RESULTS: Participants reported good task usability. Adherence to optional daily assessment was moderate. Parallel forms correlation coefficients between ASRTs were moderate-high. ASRTs correlated moderately with established tests of episodic memory and global cognitive function. Poorer performance was observed in participants with MCI/Mild AD.

DISCUSSION: Unsupervised ASRT assessment is feasible in older and cognitively impaired people. This automated task shows good parallel forms reliability and convergent validity with established cognitive tests.

1. INTRODUCTION

Now that the first disease-modifying treatment for Alzheimer's disease (AD) is available [1], there is an urgent, increased need for broader screening and improved monitoring of disease progression and treatment response in at-risk populations. Cognitive assessments are currently some of the least invasive, most cost-effective measures available.

Furthermore, they are supported for use as endpoints of treatment efficacy early in the Alzheimer's continuum by key regulatory bodies, including the US Food and Drug Administration (FDA) [2] and the European Medicines Agency (EMA) [3].

However, many cognitive assessments are lengthy, require trained personnel to administer and score and offer few parallel test variants, making them susceptible to practice effects. More importantly, test performance is measurably influenced by a range of state factors, such as sleep [4], exercise [5], mood [6] and stress [7]. This variation can give the inaccurate impression of improvement or decline over time [8]. Higher frequency sampling can generate more stable and reliable estimates of constructs of interest by controlling for state effects [9] and delineating short-term cognitive fluctuations from longer term changes associated with treatment response and disease progression [8].

Story recall is a cognitive testing paradigm used to assess verbal episodic memory and commonly used to track AD related decline [10–14]. Story recall is impaired in Alzheimer's dementia [15] and shows variable differentiation of cognitively impaired individuals with MCI from those that are cognitively unimpaired [16] and predicts progression from MCI to Alzheimer's dementia [17].

Most story recall tests are administered in person and scored manually, but research shows that scoring can be fully automated using natural language processing technologies [18]. This suggests that story recall tests could be administered in clinic at lower cost and with reduced clinician time burden. Moreover these tests may be suitable for use in remote assessment, provided that test administration can be automated.

Although remote digital assessments are not new, the SARS-CoV-2 global pandemic accelerated the need to adopt remote or hybrid clinical assessment or research methods [19,20]. Alongside advances in technology and connectivity, this has led to a growing appetite for using personal digital devices to collect clinically informative data. Beyond this, digital health technologies can enhance inclusivity, improving access for people who experience mobility problems or those with financial, geographical or time restrictions [21].

The current study describes the Automatic Story Recall Task (ASRT), a remote, self-administered and automatically scored test developed for repeated cognitive assessment, opening up opportunities for much more nuanced longitudinal data analysis. We examine test characteristics in participants who are cognitively unimpaired, have mild cognitive impairment (MCI) or have mild AD, assessed repeatedly over one week. We examine: (1) acceptability of remote ASRT assessment; (2) adherence to daily remote ASRT assessments; (3) parallel forms reliability; (4) convergent validity with established episodic memory and general cognition measures; (5) task performance characteristics; and (6) the impact of daily internal state factors.

2. METHODS

2.1 Participants

Participants were recruited from November 2020-August 2021 from three sites in the UK (London/Guildford, Plymouth, and Birmingham), and the USA (Santa Ana, California). Subjects were enrolled if they were cognitively unimpaired (CU) or diagnosed with MCI in the prior 5 years. In the UK study, participants diagnosed with mild AD in the last 5 years were

also included. MCI due to AD and mild AD diagnoses were made following National Institute of Aging-Alzheimer's Association core clinical criteria [22]. Subjects were approached if they had undergone a prior Amyloid beta (Aβ) PET scan or CSF test (confirmed Aβ- within 30 months or Aβ+ within 60 months). Eligibility was established by screening via video call using a secure Zoom link or in-clinic assessment, during which the Mini-Mental State Exam (MMSE) [23] was administered.

Inclusion criteria comprised age 50-85; MMSE score of 23-30 for participants with MCI or Mild AD, 26-30 for CU; cognitively unimpaired or clinical diagnosis of MCI/mild AD made in previous 5 years; English as a first language; availability of an informant for clinical interview (caregiver or close associate); access and ability to use a smartphone running an operation system of Android 7 or above, or iOS 11 or above.

Exclusion criteria: current diagnosis of general anxiety disorder (GAD); recent (6-month) history of unstable psychiatric illness; history of stroke within the past 2 years or a documented history of transient ischaemic attack or unexplained loss of consciousness in the last 12 months. Participants treated with medications for symptoms related to AD were required to be taking a stable dose at least 8 weeks. Participants with a current diagnosis of major depressive disorder (MDD) (UK study) or those with current or a 2-year history of MDD (US study) were excluded.

2.2 Ethics statement

This research was approved by Institutional Review Boards in the relevant research authorities (UK REC reference: 20/WM/0116; US IRB reference: 8460-JGDuffy). Informed consent was taken at the study site (US study) or electronically in accordance with HRA guidelines (UK study). Studies are registered on clinicaltrials.gov (NCT04828122, NCT04928976).

2.3 Procedure

2.3.1 Clinical assessments

Participants completed clinical assessments via a secure Zoom link (UK study) or in-clinic (US study), completed with a trained psychometrician.

The Wechsler Logical Memory Test (LMT) [24] evaluates free recall of a story according to 25 pre-defined information units (IUs: a metric quantifying the amount of information recalled, with each unit capturing the semantic essence of a part of the story [25]), immediately after presentation, and after a 30-minute delay. Paraphrased answers were accepted and scoring was completed manually using a standard scoring template.

Cognitive tests incorporated in the Preclinical Alzheimer's Cognitive Composite with semantic processing (PACC5) were administered. Tests were manually scored and a mean z-score was calculated as described in prior research [11]. The Clinical Dementia Rating scale (CDR) [26], a semi-structured interview assessing severity of cognitive symptoms of dementia, was completed with the participant and their caregiver and scored based on the CDR Sum of Boxes (CDR-SOB) scales. In the US study, where participants had completed subtests of the PACC5 or CDR assessments within one month prior to the study visit, tests were not re-administered but the recent historical test results were used.

During clinical assessments, participants were supported with installing the Novoic mobile application ('the app') on their own smartphone device and were shown how to use it.

2.3.2 Remote assessments

Participants were encouraged to complete optional unsupervised self-assessments (<30 mins) on the app daily for up to eight days following the study visit.

Remote assessments included the Automatic Story Recall Tasks (ASRTs), which are story recall tasks constructed to elicit naturalistic speech within a closed domain. ASRTs are presented at a steady reading rate (approximately 140 words per minute) by a British male speaker. Parallel stimuli available include 18 short stories (mean of 30 IUs and 119 words per story) and 18 long stories (mean of 60 IUs and 224 words per story). Task characteristics are presented in supplementary table S1, showing that stories incorporate a range of themes and are balanced for key linguistic and discourse metrics.

ASRTs were administered daily, in threes (triplets) and at the beginning of each remote assessment. Participants were asked to listen to the stories carefully. After each story was presented they were asked to immediately retell the story in as much detail as they could remember. Recall of the same stories, in the same order, was tested again after a delay (either after completion of all immediate recalls or after a distractor task). Task responses were automatically uploaded to a secure server.

Due to participant feedback regarding high burden, the assessment schedule was changed partway through the study. The new schedule favoured the use of shorter stories and reduced the number of additional assessments which followed ASRTs (not reported here). Simultaneously, the number of days of remote assessment was increased from seven to eight days to spread out assessments. Details are provided in supplementary table S2.

Daily state effects were assessed after completion of ASRTs via a four-item self-report questionnaire asking how they were feeling that day (current mood, sleep, mind-wandering and effort). App and task usability was assessed via a self-report questionnaire at the end of the assessment on day 2 (initial assessment schedule) or day 5 (revised assessment schedule). Participants reported technical difficulties experienced during assessments, whether technical difficulties prevented them from completing the assessments, how easy and how interesting it was to complete the assessments. Questionnaires are shown in supplementary tables S3 and S4.

2.4 Statistical analysis

Stories were transcribed using an out-of-the-box automatic speech recognition (ASR) system, followed by automated textual analysis completed using a generalized matching score (referred to here for brevity as "G-match"), computed in Python as the weighted sum of the cosine similarity between the embeddings of the source documentation (original ASRT text) and the transcribed retellings. G-match provides an index of the proportional recall for each story, with potential scores ranging from 0 to 1 (hypothetically perfect performance). Mean G-match per triplet was also computed.

All further analysis was completed in R v.4.0. Data were assessed for normality, followed by parametric and non-parametric analyses, as appropriate. Since a large proportion of

participants only completed seven days of remote assessments, analyses were limited to assessments on days 1-7.

Adherence to the remote testing regimen was defined as engaging with at least one ASRT story per day. Overall adherence patterns were examined with two logistic regression models, predicting adherence at immediate and delayed recall, in relation to participant group, demographic factors, assessment day and schedule. Participants were included as random effects. Demographic factors (sex, age, years in education), remote assessment days (1−7), research schedule (schedule 1 or schedule 2) and participant group (CU or MCI/mild AD) were included as fixed factors.

Parallel forms reliability of ASRTs was examined with pairwise correlational analysis. Only ASRTs administered across both schedules were analysed, to maintain comparable sample sizes across comparisons and allow for testing within MCI/mild AD and CU subgroups. Convergent validity of these same ASRT stories was examined in relation to the LMT, PACC5 and CDR-SOB. Analyses were repeated with the mean G-score per triplet. Due to variation in the distributions of tests, and to improve consistency and comparability of reporting, Spearman's rank correlation coefficients are reported.

Task performance differences between groups, task administration variations and over time were modeled using longitudinal linear mixed effects models. This included G-match for individual ASRTs as the response variable, and fixed effects of participant group, remote assessment days (1-7), order (1st, 2nd or 3rd ASRT presented), long or short stories and immediate or delayed recall. Demographic covariates (age, sex, education) were included as additional fixed effects. A random effect of participant with random slope and intercept was specified.

Analyses were repeated with the mean G-match per triplet, with equivalent random and fixed effects specifications, excepting story order which was not included. Covariation of mean ASRT task performance across triplets with self-reported daily state was then examined, by additionally incorporating fixed effects of self-reported mood, sleep, effort and mind-wandering, into the above model. Assumptions of regression models were investigated by examining the distribution and patterns of residuals versus fitted values.

3. RESULTS

3.1 Participants

Two hundred participants, 67 from the US study and 133 from the UK study, were recruited. One hundred and fifty-one participants (75.5%) completed at least one remote ASRT. Older participants (r=-0.15, p= 0.03), those with higher MMSE scores (r=-0.26, p< 0.001) and those with MCI/mild AD (73/106 MCI (69%) and 78/94 CU (83%); χ^2=5.36 (DF=1), p=0.02) were less likely to complete any remote assessments. There were no differences in sex ratio (χ^2=0.41 (DF=1), p=0.52), or years in education (r=-0.01 p-value = 0.87) between participants who contributed at least one remote assessment and those who did not.

Demographic information for the 151 participants providing remote data are presented in table 1. In this subsample, MCI and CU groups did not differ with respect to key demographic factors (age, years in education, sex) or amyloid status. Proportionally more participants with MCI were recruited and completed remote assessments in the US study.

Group	**Number of participants from UK/ US study**	**Number of participants (female/ male)**	**Amyloid negative/ positive**	**Schedule 1 versus schedule 2**	**Mean years in education (SD)**	**Mean age (SD)**	**Mean MMSE (SD)**
Cognitively unimpaired	66/12	78 (47/31)	38/40	40/38	15.24 (3.37)	70.37 (4.35)	28.92 (1.15)
MCI/mild AD	51/22	73 (41/32)	41/32	22/51	15.06 (2.97)	69.58 (7.30)	27.00 (2.06)
Statistic and p-value	χ^2=4.70 p<0.03	χ^2=0.48 p=0.49	χ^2=0.16 p=0.69	χ^2=11.89 p<0.001	r=-0.08, p= 0.30	r=-0.08, p= 0.35	r=0.58, p<0.001

Table 1: Participant demographic characteristics: Demographic characteristics of cognitively unimpaired and MCI/mild AD participants. MCI: mild cognitive impairment; AD: Alzheimer's dementia; CU: cognitively unimpaired; N, number; SD, standard deviation. MMSE: Mini-Mental State Exam.

3.2 Usability

Usability questionnaires were completed by 96 participants (n=52 CU, n=44 MCI/mild AD), with results presented in figure 1. Participants reported few technical difficulties and most reported that technical difficulties had not prevented them from completing the assessments, with no group differences (χ^2=3.32 (DF=1), p=0.07 and χ^2=0.98 (DF=1), p=0.32, respectively). Participants overwhelmingly responded that the app was easy to use, and that the task was reasonably interesting, again with no group differences (r=-0.08, p=0.47 and r=-0.04, p=0.70, respectively).

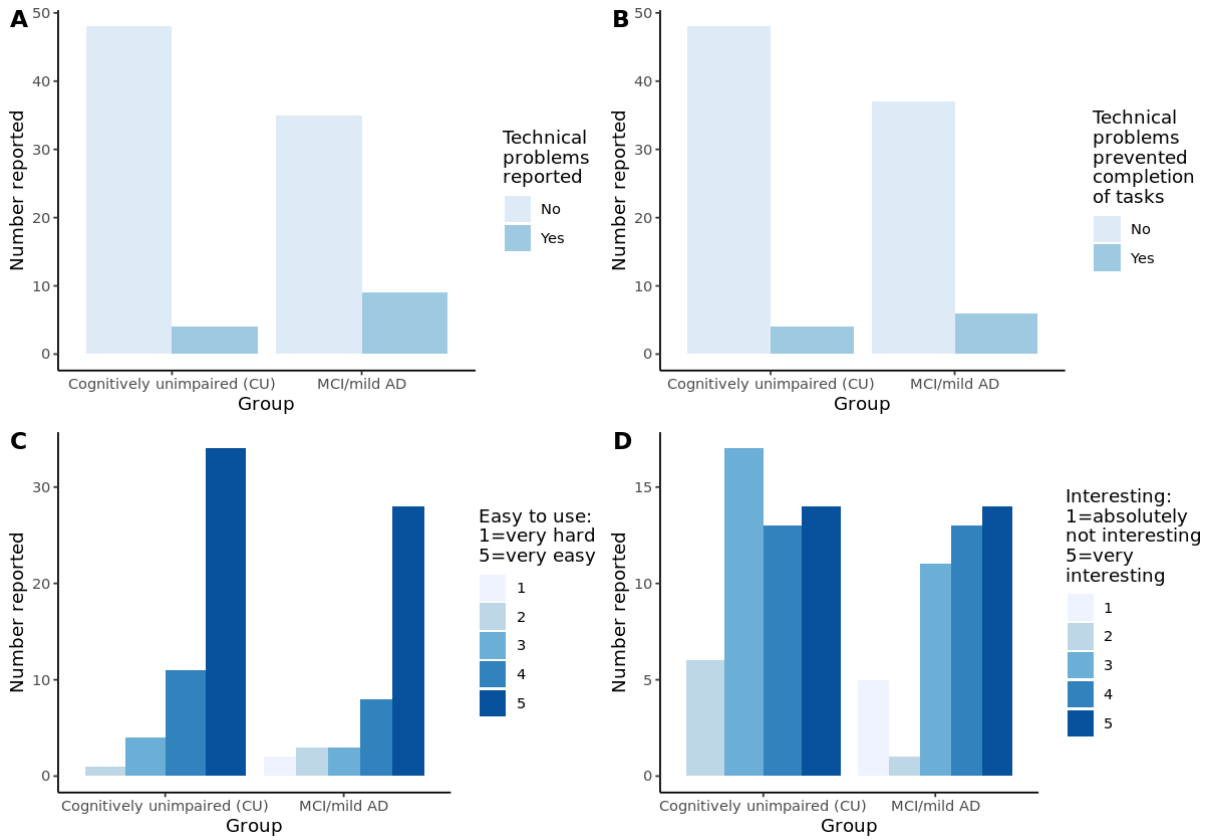

Figure 1: responses to usability questionnaire: A) technical problems reported; B) rate at which technical problems prevented completion of tasks; C) ease of use of application; D) interest in tasks completed.

3.3 Adherence

Participants with MCI/mild AD completed fewer remote assessments than CU participants (adherence for immediate recall: 66% versus 78%; delayed recall: 63% versus 77%; figure 2). Group differences in adherence were confirmed by mixed logistic regression analyses (immediate recall, estimate=-1.00, p=0.01; delayed recall estimate =-0.84, p=0.02). Adherence did not change over time (immediate recall estimate=-0.04, p=0.30; delayed recall estimate=-0.07, p=0.13), but lower adherence to delayed recall was seen for the revised test schedule (fixed effects estimate=-0.86, p=0.03). Adherence was not associated with sex and education (all p>0.2), but younger participants completed more assessments (fixed effects estimates: immediate=-0.07, p=0.03; delayed=-0.06, p=0.06).

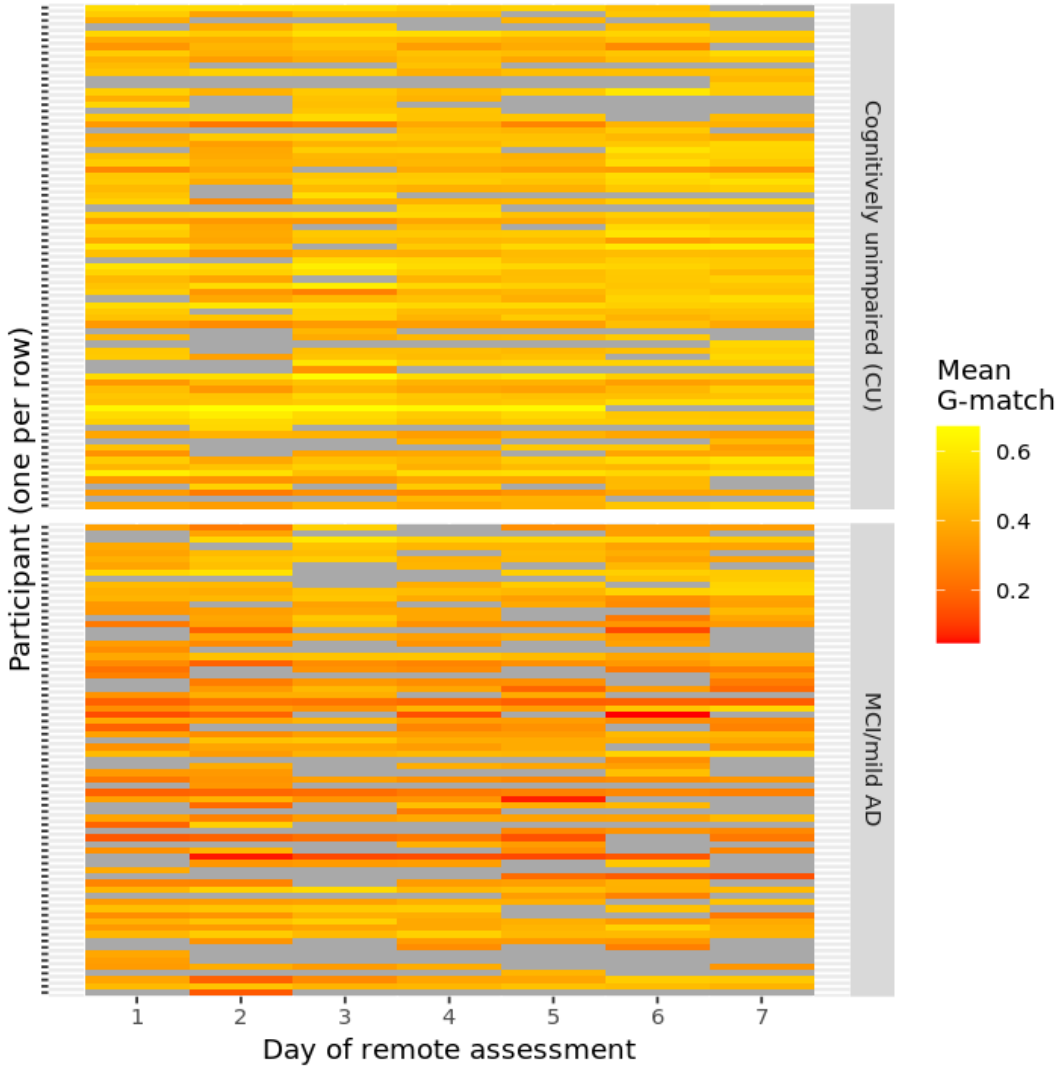

Figure 2: Adherence and task performance heatmap for G-match in immediate recall trials. G-match is an automated measure of recall performance (see methods). Results plotted across individual days of remote assessment for n=151 participants who completed at least one assessment. Each participant is represented by a row, missing data are shown in grey, and mean G-match across ASRT triplets is shown in colour (red=low recall; yellow=high recall).

3.4 Task characteristics

G-match for ASRTs and triplets showed good psychometric properties, with no ceiling or floor effects (supplementary Figure SF1-4). Task performance characteristics separated by group and immediate and delayed recall, are provided in supplementary tables S5-S7. Longer stories elicited a greater number of spoken words, but G-match for longer stories was typically lower, indicating that high levels of recall similarity to original text are harder to achieve with longer source texts. Figure 3A provides an overview of recall for individual ASRTs, showing variability between parallel ASRTs and long and short stories.

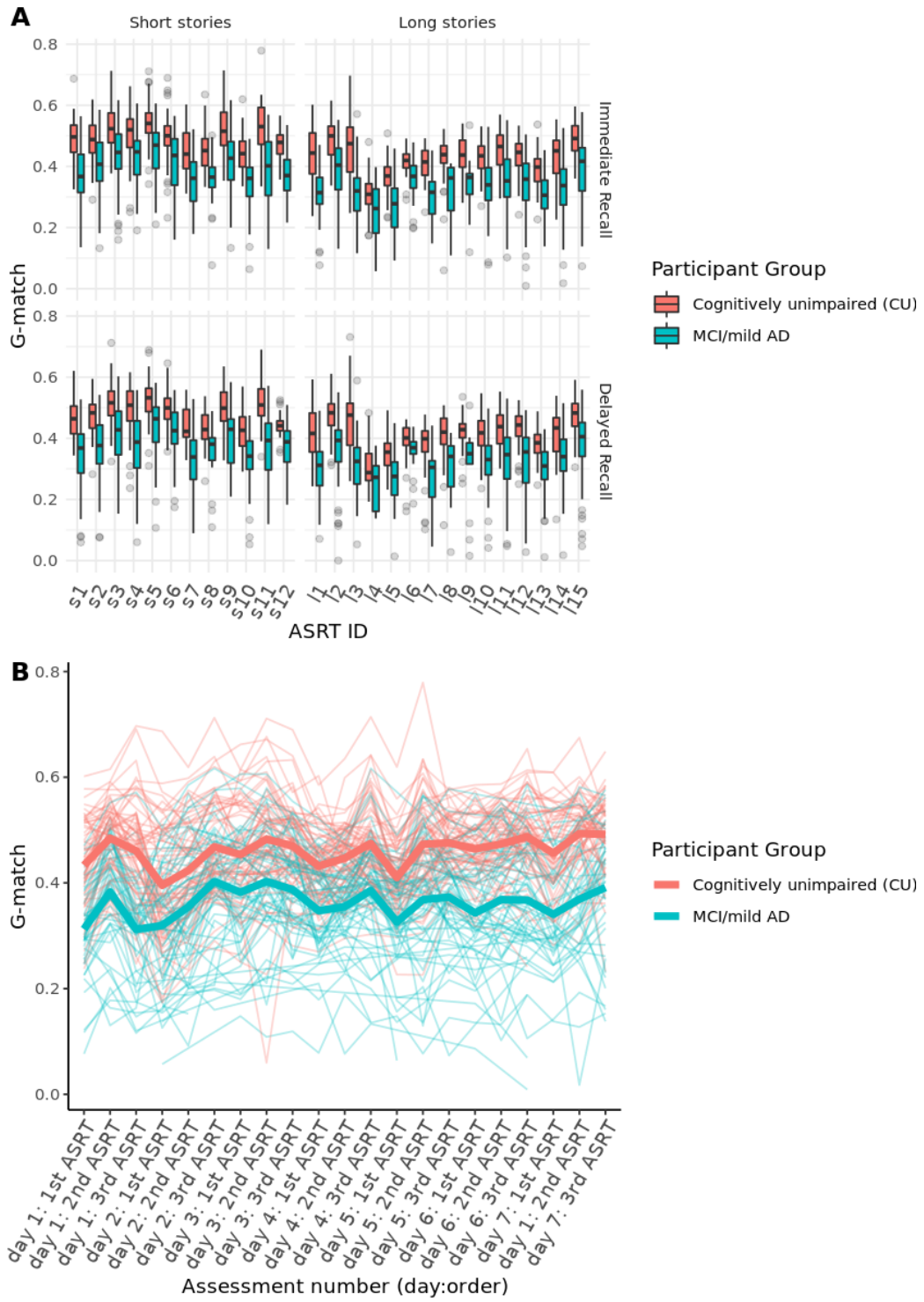

Figure 3: G-match, an automated measure of recall performance, over repeated assessments: A) boxplots of G-match in individual ASRT stories split by short and long story horizontally and by immediate and delayed recall vertically. Participant groups (cognitively unimpaired, MCI/mild AD) plotted in different colours; B) group means for immediate recall in thick lines and individual variability across remote assessment days and testing order with individual trajectories in the paler, thinner lines, showing variability within individuals, across testing days and order of administration.

3.5 Parallel form reliability

Results show good reliability across the parallel ASRTs (figure 4). Correlation coefficients between ASRT stories in the full sample were moderate to strong for immediate recall (rho range=0.56-0.88, mean=0.72), and remained so after restricting analyses to MCI/Mild AD (rho range=0.31-0.86, mean=0.64) and CU (rho range 0.39-0.85, mean=0.64). Similarly, correlations between parallel ASRT stories were high for delayed recall (full sample: rho range=0.56-0.86, mean=0.74), and remained so when restricting analyses to MCI/mild AD (rho range=0.37-0.88, mean=0.65) and CU participants (rho range=0.32-0.83, mean=0.64).

Test-retest reliability was even higher when examined for mean scores obtained across triplets (immediate; rho range=0.77-0.88, mean=0.83; delayed: rho range=0.84-0.89, mean=0.86), remaining consistently high in MCI (immediate; rho range=0.64-0.88, mean=0.75; delayed: rho range=0.68-0.84, mean=0.77) and CU subgroups (immediate; rho range=0.67-0.83, mean=0.76; delayed: rho range=0.73-0.85, mean=0.79).

Parallel forms reliability correlation coefficients for G-match for individual ASRT stories (immediate recall) are presented in figure 4. Equivalent figures for delayed recall, and additionally separated by clinical group, are presented in supplementary Figures SF 5-9. Correlation matrices for triplets broken down by immediate and delayed, and clinical groups, are shown in supplementary Figures SF 10-12.

3.6 Convergent validity

ASRT task performance correlated moderately with other cognitive measures (LMT, CDR-SOB and PACC5) in the full sample across both immediate and delayed recalls. Convergent validity for immediate recall ASRTs is shown in figure 4. Delayed recall, and correlation coefficients separated by participant group are provided in supplementary Figures SF 5-9). Correlation coefficients remained in the moderate range after restricting analyses to participants with MCI/mild AD but were typically lower in CU participants.

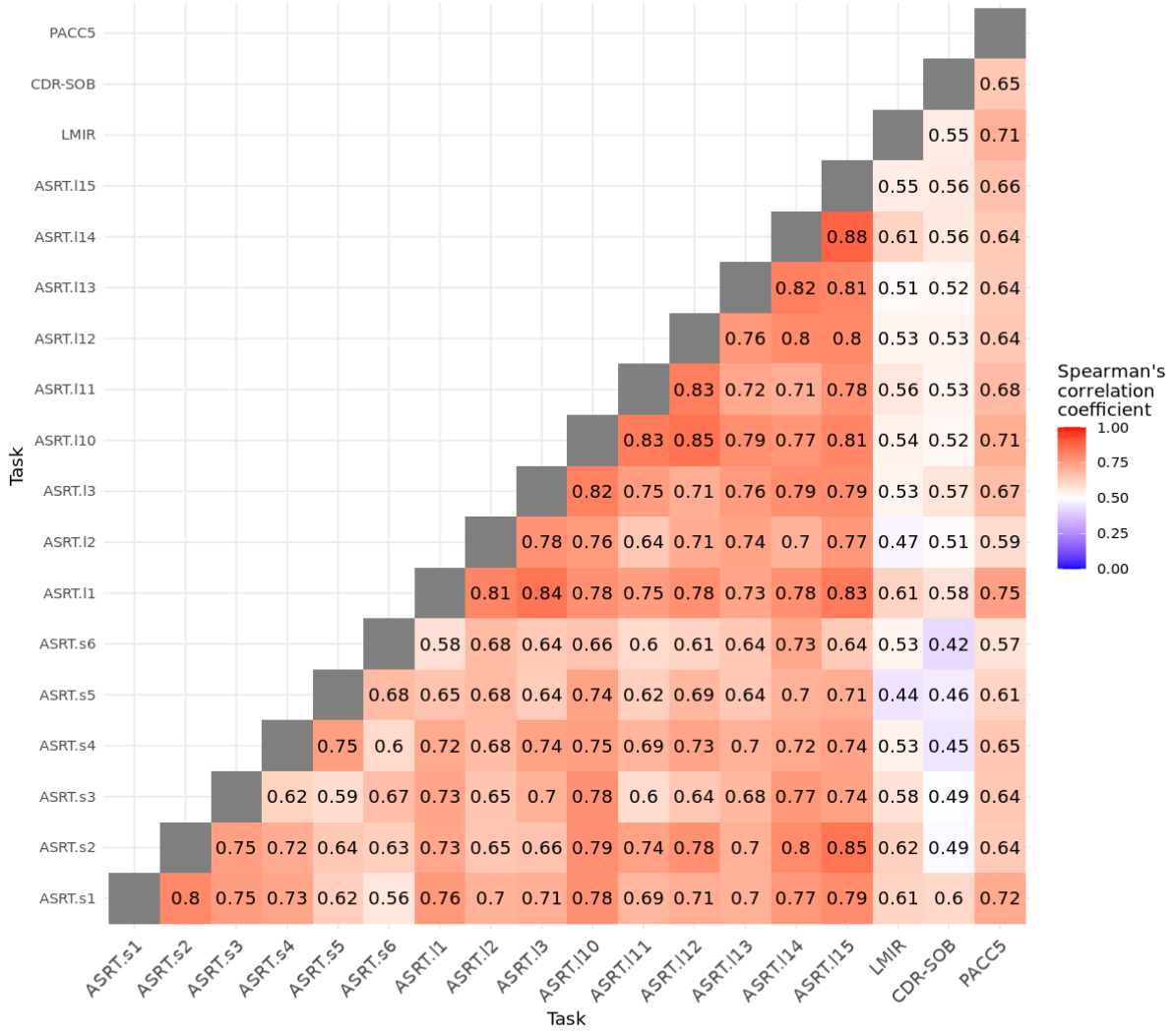

Figure 4: Matrix of pairwise correlation coefficients of test-retest reliability for G-match, an automated measure of recall performance, of individual ASRT stories. Convergent validity is examined through correlations between ASRT stories with other cognitive test scores (LMIR, CDR-SOB, PACC5) obtained during clinical assessments. To maintain consistent reporting, the sign of the correlation for the CDR-SOB is reversed in the above figure, meaning that for all tests higher scores denote better cognitive test performance. Pairwise correlation coefficients for ASRTs reflect parallel task performance metrics for between n=80-116 participants, and with other cognitive tests for n=89-149, depending on adherence patterns. Abbreviations: ASRT: Automatic Story Recall Task, LMIR: Wechsler Logical Memory Test immediate recall, CDR-SOB: Clinical Dementia Rating - Sum of Boxes, PACC5: Preclinical Alzheimer's Cognitive Composite with semantic processing.

3.7 Task performance

Longitudinal mixed models of task performance by study day are presented in table 2, revealing similar results for individual ASRTs and triplets. Task performance modestly improved across the week. There was an effect of group, with lower scores in the MCI/mild

AD group. G-match was higher for immediate recall, shorter stories, and more elevated for the latter ASRTs administered within each triplet. Demographic measures were not associated with task performance. Results are displayed in figure 3B, showing within and between-subject variability.

After incorporating self-report assessments into the mixed model predicting average G-match for triplets, models revealed a significant effect of mood (estimate=0.006 (SE=0.001), p<0.001) and mind-wandering (estimate=-0.009 (SE=0.001), p<0.001) on daily ASRT performance, with better daily mood and lower mind-wandering associated with better task daily performance.

Response variable	**ASRTs: model parameters**			
	Predictors	**Estimate**	**SE**	**p-value**
G-match (individual stories)	Intercept	0.56	0.08	<0.001
	Group (cognitively unimpaired, MCI/mild AD)	-0.11	0.01	<0.001
	Assessment day	0.002	0.001	0.02
	Recall type (immediate versus delayed)	-0.02	0.002	<0.001
	ASRT length (long/short)	-0.05	0.002	<0.001
	ASRT order of presentation (1,2,3)	0.02	0.001	<0.001
	Sex (male/female)	-0.02	0.01	0.09
	Education	0.0001	0.002	0.95
	Age	-0.002	0.001	0.09
G-match (triplets)	Intercept	0.60	0.08	<0.001
	Group (cognitively unimpaired, MCI/mild AD)	-0.12	0.01	<0.001
	Assessment day (1-7)	0.002	0.001	0.02
	Recall type (immediate/delayed)	-0.02	0.02	<0.001
	ASRT length (long/short)	-0.05	0.003	<0.001
	Sex (male/female)	-0.02	0.01	0.08
	Education	0.0002	0.002	0.93
	Age	-0.002	0.001	0.08

Table 2: Effects of task characteristics, participant group, and demographics on task performance metrics as estimated by longitudinal mixed models.

4. DISCUSSION

Older participants, with and without cognitive impairment, engaged in optional daily remote unsupervised speech assessments with moderate levels of adherence. Task performance

on the ASRT differed between clinical groups. Subjects experienced few technical problems and reported that the tests were easy to use and reasonably interesting. Results show that remote automatic test administration and auto-scoring of test performance is feasible and can provide sensitive cognitive measurement in key populations.

ASRT stimuli were carefully designed and balanced, which in turn is reflected in good parallel forms reliability, with moderate to high correlations between ASRT variants. ASRT tests also correlated moderately with a well established test of verbal episodic memory and tests global cognition, indicating acceptable convergent validity, and with results comparable to, or better than, other studies of computerised or unsupervised remote assessments [27–29]. The current study also examined correlation coefficients within the two clinical groups. Convergent validity was typically lower in CU participants, which can be linked to ceiling or floor-level performance on certain traditional cognitive tests in healthy individuals.

The lack of decline in adherence over time indicates that testing over a weeklong period does not produce testing fatigue effects sufficient to have an impact on rates of daily participation. Nor does continued participation adversely affect test scores, since these improved modestly during the week, indicating that increased familiarity with the app, testing procedure and/or test structure resulted in a subtle improvement over time.

The current study shows within-subject variation in task performance, in part reflecting measurable effects of state factors on cognitive performance, in particular daily mood and effort. Task performance differences also reflect aspects of study design, with stories administered later in the triplets delivering more comprehensive recall than those administered earlier, and longer stories producing less comprehensive recall than shorter stories.

4.1 Limitations

A higher proportion of older and more cognitively impaired participants did not engage in remote assessment. This may reflect that remote assessments themselves were optional. Although most participants were able to engage with remote assessment, the testing schedule was altered in the middle of the study to reduce participant burden, thereby limiting the amount of data available for certain ASRT parallel test variants. Overall these findings indicate that brief remote assessments are likely to be more acceptable in this population. Assessment under supervision, either in clinic or during a telemedicine visit, could be more appropriate for more impaired subjects.

The design of the study makes it difficult to differentiate between the effects of individual stories themselves (i.e. which ASRT story was used) and effects of study design, such as test order or day of assessment. Future studies may benefit from adopting a randomised design, with ASRTs randomly selected and allocated to different testing instances, to derive test performance metrics independent of these additional confounders. For longitudinal studies, either short or long stories should be adopted to improve consistency of test scores over time and help to better characterise cognitive change.

4.2 Overview and future directions

The recent FDA approval for the first disease modifying treatment for people at risk of developing AD highlights the importance of adequate screening and early detection, as well

as the importance of monitoring treatment response. Briefer, convenient and lower-burden daily assessments may provide more reliable data to evaluate disease progression or treatment response than one-off lengthy assessments [9]. The current study shows that brief, remotely administered and automatically scored ASRTs are sensitive to early cognitive impairments commonly identified through more extensive clinical assessment. The tests show good properties for repeated administration, and convergent validity with established tests of episodic memory and global cognitive function.

Speech responses are a common component of cognitive tests, however data generated in these tests, including those reported in this study, often relate to simple pass/fail characteristics of response accuracy. New metrics using audio- and text-based AI models to target other changes measurable in speech data (acoustic [30,31], semantic [32–35], linguistic [31]) in early-stage Alzheimer's disease could further leverage the information content of ASRTs, developing a new class of powerful, fully automated speech biomarkers.

Acknowledgements/Conflicts/Funding Sources

The study was funded by Novoic, a clinical stage digital medtech company developing AI-based speech biomarkers. The study funder was involved in study design, data interpretation and writing of the report.

EF, JW, MM, CS, RL and UM are employees of Novoic Ltd. KVP is an advisor to the company. EF, JW and MM are shareholders and CS, MM, RL, UM, and KVP are option holders in the company. KVP has served as a paid consultant for Biogen Idec and Digital Cognition Technologies.

Key Words

Episodic memory; speech; psychometrics; reliability; validity; ageing; Alzheimer’s disease; mild cognitive impairment; digital health; longitudinal.

References

[1] U.S Food and Drug Administration. FDA Grants Accelerated Approval for Alzheimer’s Drug 2021. https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-alzheimers-drug (accessed June 29, 2021).

[2] Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer’s disease. N Engl J Med 2013;368:1169–71. https://doi.org/10.1056/NEJMp1302513.

[3] European Medicines Agency. Guideline on the clinical investigation of medicines for the treatment of Alzheimer’s disease 2018. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-investigation-medicines-treatment-alzheimers-disease-revision-2_en.pdf (accessed April 12, 2021).

[4] Lo JC, Groeger JA, Cheng GH, Dijk D-J, Chee MWL. Self-reported sleep duration and cognitive performance in older adults: a systematic review and meta-analysis. Sleep Med 2016;17:87–98. https://doi.org/10.1016/j.sleep.2015.08.021.

[5] Chang YK, Labban JD, Gapin JI, Etnier JL. The effects of acute exercise on cognitive

performance: a meta-analysis. Brain Res 2012;1453:87–101. https://doi.org/10.1016/j.brainres.2012.02.068.

[6] Mitchell RLC, Phillips LH. The psychological, neurochemical and functional neuroanatomical mediators of the effects of positive and negative mood on executive functions. Neuropsychologia 2007;45:617–29. https://doi.org/10.1016/j.neuropsychologia.2006.06.030.

[7] Angelidis A, Solis E, Lautenbach F, van der Does W, Putman P. I'm going to fail! Acute cognitive performance anxiety increases threat-interference and impairs WM performance. PLoS ONE 2019;14:e0210824. https://doi.org/10.1371/journal.pone.0210824.

[8] Meier IB, Buegler M, Harms R, Seixas A, Çöltekin A, Tarnanas I. Using a Digital Neuro Signature to measure longitudinal individual-level change in Alzheimer's disease: the Altoida large cohort study. Npj Digital Med 2021;4:101. https://doi.org/10.1038/s41746-021-00470-z.

[9] Schweitzer P, Husky M, Allard M, Amieva H, Pérès K, Foubert-Samier A, et al. Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. Int J Methods Psychiatr Res 2016. https://doi.org/10.1002/mpr.1521.

[10] Donohue MC, Sperling RA, Salmon DP, Rentz DM, Raman R, Thomas RG, et al. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. JAMA Neurol 2014;71:961–70. https://doi.org/10.1001/jamaneurol.2014.803.

[11] Papp KV, Rentz DM, Orlovsky I, Sperling RA, Mormino EC. Optimizing the preclinical Alzheimer's cognitive composite with semantic processing: The PACC5. Alzheimers Dement (N Y) 2017;3:668–77. https://doi.org/10.1016/j.trci.2017.10.004.

[12] Insel PS, Weiner M, Mackin RS, Mormino E, Lim YY, Stomrud E, et al. Determining clinically meaningful decline in preclinical Alzheimer disease. Neurology 2019;93:e322–33. https://doi.org/10.1212/WNL.0000000000007831.

[13] Lim YY, Snyder PJ, Pietrzak RH, Ukiqi A, Villemagne VL, Ames D, et al. Sensitivity of composite scores to amyloid burden in preclinical Alzheimer's disease: Introducing the Z-scores of Attention, Verbal fluency, and Episodic memory for Nondemented older adults composite score. Alzheimers Dement (Amst) 2016;2:19–26. https://doi.org/10.1016/j.dadm.2015.11.003.

[14] Jonaitis EM, Koscik RL, Clark LR, Ma Y, Betthauser TJ, Berman SE, et al. Measuring longitudinal cognition: Individual tests versus composites. Alzheimers Dement (Amst) 2019;11:74–84. https://doi.org/10.1016/j.dadm.2018.11.006.

[15] Porto MF, Benitez-Agudelo JC, Aguirre-Acevedo DC, Barceló-Martinez E, Allegri RF. Diagnostic accuracy of the UDS 3.0 neuropsychological battery in a cohort with Alzheimer's disease in Colombia. Appl Neuropsychol Adult 2021:1–9. https://doi.org/10.1080/23279095.2021.1897007.

[16] Chapman KR, Bing-Canar H, Alosco ML, Steinberg EG, Martin B, Chaisson C, et al. Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. Alzheimers Res Ther 2016;8:9. https://doi.org/10.1186/s13195-016-0176-z.

[17] Belleville S, Fouquet C, Hudon C, Zomahoun HTV, Croteau J, Consortium for the Early Identification of Alzheimer's disease-Quebec. Neuropsychological Measures that Predict Progression from Mild Cognitive Impairment to Alzheimer's type dementia in Older Adults: a Systematic Review and Meta-Analysis. Neuropsychol Rev 2017;27:328–53. https://doi.org/10.1007/s11065-017-9361-5.

[18] Lehr M, Prud'hommeaux E, Shafran I, Roark B. Fully Automated Neuropsychological

Assessment for Detecting Mild Cognitive Impairment . INTERSPEECH 2012 2012.

[19] Ferrar J, Griffith GJ, Skirrow C, Cashdollar N, Taptiklis N, Dobson J, et al. Developing digital tools for remote clinical research: how to evaluate the validity and practicality of active assessments in field settings. J Med Internet Res 2021;23:e26004. https://doi.org/10.2196/26004.

[20] Ousset PJ, Vellas B. Viewpoint: Impact of the Covid-19 Outbreak on the Clinical and Research Activities of Memory Clinics: An Alzheimer's Disease Center Facing the Covid-19 Crisis. J Prev Alzheimers Dis 2020;7:197–8. https://doi.org/10.14283/jpad.2020.17.

[21] National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Forum on Drug Discovery, Development, and Translation. Virtual clinical trials: challenges and opportunities: proceedings of a workshop. Washington (DC): National Academies Press (US); 2019. https://doi.org/10.17226/25502.

[22] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9. https://doi.org/10.1016/j.jalz.2011.03.008.

[23] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state" A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–98.

[24] Wechsler D, Stone CP. Wechsler Memory Scale-Revised. San Antonio, TX: Psychological Corporation; 1987.

[25] McNeil MR, Doyle PJ, Fossett TRD, Park GH, Goda AJ. Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. Aphasiology 2001;15:991–1006. https://doi.org/10.1080/02687040143000348.

[26] Morris JC. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology 1993;43:2412–4. https://doi.org/10.1212/wnl.43.11.2412-a.

[27] Papp KV, Samaroo A, Chou H, Buckley R, Schneider OR, Hsieh S, et al. Unsupervised mobile cognitive testing for use in preclinical Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 2021;13. https://doi.org/10.1002/dad2.12243.

[28] Mackin RS, Rhodes E, Insel PS, Nosheny R, Finley S, Ashford M, et al. Reliability and Validity of a Home-Based Self-Administered Computerized Test of Learning and Memory Using Speech Recognition. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2021:1–15. https://doi.org/10.1080/13825585.2021.1927961.

[29] Busch RM, Hogue O, Ferguson L, Parsons MW, Kubu CS, Floden DF. Validation of computerized episodic memory measures in a diverse clinical sample referred for neuropsychological assessment. Clin Psychol 2019;3:557–70.

[30] Martínez-Nicolás I, Llorente TE, Martínez-Sánchez F, Meilán JJG. Ten years of research on automatic voice and speech analysis of people with alzheimer's disease and mild cognitive impairment: A systematic review article. Front Psychol 2021;12:620251. https://doi.org/10.3389/fpsyg.2021.620251.

[31] Roark B, Mitchell M, Hosom J-P, Hollingshead K, Kaye J. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans Audio Speech Lang Processing 2011;19:2081–90. https://doi.org/10.1109/TASL.2011.2112351.

[32] Foldi NS. Getting the hang of it: preferential gist over verbatim story recall and the

roles of attentional capacity and the episodic buffer in Alzheimer disease. J Int Neuropsychol Soc 2011;17:69–79. https://doi.org/10.1017/S1355617710001165.

[33] Mueller KD, Koscik RL, Du L, Bruno D, Jonaitis EM, Koscik AZ, et al. Proper names from story recall are associated with beta-amyloid in cognitively unimpaired adults at risk for Alzheimer's disease. Cortex 2020;131:137–50. https://doi.org/10.1016/j.cortex.2020.07.008.

[34] Drummond C, Coutinho G, Fonseca RP, Assunção N, Teldeschi A, de Oliveira-Souza R, et al. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. Front Aging Neurosci 2015;7:96. https://doi.org/10.3389/fnagi.2015.00096.

[35] Bruno D, Mueller KD, Betthauser T, Chin N, Engelman CD, Christian B, et al. Serial position effects in the Logical Memory Test: Loss of primacy predicts amyloid positivity. J Neuropsychol 2020. https://doi.org/10.1111/jnp.12235.