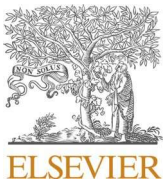
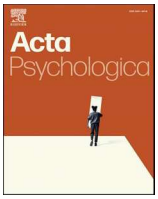
Acta Psychologica 257 (2025) 105084



Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



AI-powered speech device as a tool for neuropsychological assessment of an older adult population: A preliminary study

Daniela E. Aguilar Ramirez [a,*], Lukas Grasse [a,b], Scott Stone [a,b], Matthew Tata [a,b], Claudia L.R. Gonzalez [a]

[a] Department of Kinesiology and Physical Education, University of Lethbridge, 4401 University Drive, Lethbridge, AB, Canada
[b] Reverb Robotics Inc., Lethbridge, AB, Canada

ARTICLE INFO

Keywords:
Aging population
Artificial intelligence (AI)
Healthcare systems
Memory tasks
Neuropsychological tests

ABSTRACT

As the older adult population continues to expand, the demands on the healthcare system intensifies, necessitating the development of technologies that effectively accommodate the requirements of older adults. While Artificial Intelligence (AI) systems hold promise as a solution, they have not been designed to accommodate the sensory and cognitive changes typical of aging individuals. The current study investigates the use of an AI-powered communication device for the assessment of neuropsychological tests to an older adult population. Twenty-four (twelve females) older adult participants completed three memory tasks using the AI device: logical memory, poem recall, and the backward and sequencing digit span tests. Significant negative correlations were found between the age of the participants and performance on the Logical memory and digit span tests. The AI device effectively identified age-related memory changes comparable to those observed with human administrators. Implementing this technology in healthcare offers several advantages: alleviating healthcare professionals' workload, improving standard of care by reaching underserved populations, and facilitating continuous screening for early identification of prodromal stages of neurodegenerative diseases.

1. Introduction

Aging, caused by the accumulation of damage in cells over time, is a natural and progressive process that occurs in every human being. With senescence, physiological changes in all organ systems emerge. Over the course of a lifetime, organ function begins to decline, leading to a cascade of age-related diseases and conditions. The risks for neurodegenerative conditions and diseases significantly increases with age and many older adults live with multiple comorbidities (Guo et al., 2022). Currently, societies across the world are facing a looming health crisis with an increasing aging population, mostly attributed to the "baby boomer" generation. By 2030, one in six people in the world will be 60 years or older (World Health Organization, 2022), and in Canada and the US this number will be even larger with one in four and one in five respectively (US Census Bureau, 2018). Facilities for long-term senior care, more commonly used in North America, soon will not be capable of accommodating for this aging population. Two important repercussions are of concern. First, a growing cohort of older adults will not have access to the limited assisted living options thus, increasing the time that this population will be living independently. Second, the health care workforce will encounter unprecedented strain due to escalating care demands. One solution that can alleviate these concerns is the integration of artificial intelligence (AI) systems into the everyday lives of older adults and their caregivers. In general, AI will allow for remote accessibility, enabling better quality of life for older adults living at home. Furthermore, these devices can help augment the abilities of health care professionals; as a result, workloads can be decreased while increasing productivity.

Advances in AI have led to 'superhuman-level' performance in different tasks. This was made clear over two decades ago, in 1997, when IBM's Deep Blue computer defeated the world's top-ranked chess player, Garry Kasparov. Significant changes have taken place since the era of Deep Blue. Most notably the computer power, the amount of data that is available for training and analyses, deep learning enabling models to learn hierarchical representations of data, and cloud computing (Dean, 2022). Although AI is still far from replicating the complexity of the human brain, AI has been successful at completing distinct tasks at a level outperforming humans (Mnih et al., 2013, 2015).

* Corresponding author.
E-mail address: d.aguilarramirez@uleth.ca (D.E. Aguilar Ramirez).

https://doi.org/10.1016/j.actpsy.2025.105084
Received 24 February 2025; Received in revised form 2 May 2025; Accepted 15 May 2025

For example, billions of people around the world use face and voice recognition systems and benefit from the use of Google's search engine, Google maps, Google Lens, Google assistant or Google Home, and (more currently) ChatGPT, together with a vast amount of other AI powered systems. Although less notoriously but more importantly, AI systems have revolutionized scientific research, medical imaging, and healthcare (Yu et al., 2021; Sirilertmekasakul, Rattanawong, Gongvatana, & Srikiatkhachorn, 2023). Because of AI's capacity to gather, process, and analyze large amounts of data, it can produce meaningful information and form reliable predictions. In healthcare, AI has not only improved workflow in healthcare facilities, but it has been particularly valuable at detecting meaningful changes in a patient's health at early stages, achieving high diagnostic precision, and reducing false positives (Ardila et al., 2019; Kaiser, Hitzl, & Iglseder, 2014; McKinney et al., 2020; Yu et al., 2021). In research, neuroscientists have used AI to gain deeper understanding into the human brain and to enhance the accuracy of diagnoses for neurodegenerative conditions such as Parkinson's (Park, Kim, Kim, & Lee, 2020) and Alzheimer's disease (Chen et al., 2020). Overall, AI has accelerated the pace of scientific research and healthcare delivery while simultaneously improving precision.

Unfortunately, even when societies are facing a crisis with a rapid growth of an older population, the technological industry has placed little attention to gerontechnology (defined as software and devices built with the focus to fulfill the needs of older people; Chen, 2020). The currently available AIs are not suitable for older adult populations as they are not user friendly, nor designed for a population in which sensory abilities (i.e. vision, audition, and touch) decline with an increase in age (Völter et al., 2021). In 2022, the WHO warned about the risks of ageism in AI (World Health Organization, 2022). They stated that ageism may be encoded in the data of AI powered systems. For example, it is likely that some AI devices are trained on data derived from younger populations due to a younger generation that is familiar with technological devices. Training, validation, implementation, and assessment of these systems in older adults is needed. The WHO further remarks on the fact that speech recognition systems have left older adults at a disadvantage, as AIs are optimized with speech patterns of the average young adult. In fact, AIs for speech recognition are less accurate at identifying older-adult voices when compared to younger populations (Kwon et al., 2016). Older adults appear to have a different notion of an AI device, they do not see the benefit or value in them (Trajkova & Martin- Hammond, 2020), and as a result less speech data has been gathered from an older adult population. Furthermore, AI devices must account for the differences in speech patterns that appear with aging. Older adults tend to have slower: articulation speed (Lee, 2011), speech rate (Lee, 2011), fainter pitch (Kim, 2003), and poorer speech intelligibility (Kwon et al., 2016). Moreover, in an older population there is the prevalence of neurodegenerative diseases such as Parkinson's in which speech impediments such as hypophonia and dysarthria are common. Speech recognition systems must also account for hearing loss, as older populations hearing sensitivity diminishes (Glyde, Hickson, Cameron, & Dillon, 2011). It is critical, therefore, to develop speech recognition devices that account for these sensory deficits found commonly in an older population (Pacala & Yueh, 2012). The use of speech devices in aging research and in health care in general should be versatile to include simple tasks, such as medication reminders, or complex ones such as administrating cognitive or neuropsychological tests. Such a device could help with the current and looming health care crisis. It will reduce workload and avoid burnout of health care providers.

In this study, we assessed an AI-powered speech device: Audio Board for Robotics and Automation (ABRA; Reverb Robotics, Lethbridge, Alberta, Canada) as a tool for research and health care in an older population. Despite the expected limitations we would face when collecting data from an older population (i.e. lower accuracy for speech-to-text, difficulty for the user to hear the output, and various other characteristics about older adults' speech that may prove challenging for the AI device to understand), we chose to use state-of-the-art artificial neural networks for speech understanding and speech output to allow two-way voice communication between the human and the system. Specifically, ABRA was used for the assessment of commonly used neuropsychological memory tests (the Weschler's memory scale and the digit span test). Understanding the shortcomings of these state-of-the-art models will help inform how to better improve them when targeting a specific population such as older adults. We hypothesized results from the neuropsychological tests administered by ABRA would be reliable enough such that it could be used to aid the human health care worker in data collection.

2. Methods

2.1. Participants

Twenty-four older adult participants (12 females and 12 males) between the ages of 63 to 78 (M = 70.6, SD = 4.7, Mdn =71) took part in the study. A priori power analysis was conducted using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) to determine the required sample size. Although power calculations were performed for multiple effects observed in our preliminary investigation, our target sample size was based on the analysis requiring $N = 23$ participants. This calculation was derived from the significant correlation found between participant age and Logical Memory delayed recall scores ($r = 0.53$). This specific analysis was chosen as the primary basis for the sample size determination because it represents a key preliminary finding related to age effects, and the required N of 23 is slightly more conservative than calculations based on other significant effects observed (e.g., $N = 19$ for age/digit span correlation; $N = 8$ for immediate/delayed unit scores). A correlation analysis ($\alpha = 0.05$, Power = 0.80, two-tailed) indicated a minimum $N = 23$. Participants lived independently, were able to give informed consent, and to complete all testing. They were also absent of speech or hearing impairments, and without any diagnosed neurological conditions. Educational attainment was evenly represented among participants, with comparable numbers across high school ($n = 6$), college($n = 7$), bachelor's ($n = 5$), and graduate (n = 6) levels. Furthermore, all participants were Caucasian, with 79 % of them being retired and 83 % married. Recruitment solely happened through word-of-mouth around the Lethbridge community primarily via the experimenter's personal and professional connections. Some of the participants had previously taken part in other research studies at the University of Lethbridge, indicating familiarity with the research process and a willingness to participate in cognitive testing. The experiment was approved by the University of Alberta Research Ethics Board 2 through the Alberta Research Information Services System.

2.2. AI system-hardware and software

All memory tests consisted of speech interactions between the participant and the ABRA. The following sections outline the hardware and software used to develop the speech interaction.

2.2.1. Hardware

The speech interaction device utilized in this study was an end-to-end hardware and software platform supplied by Reverb Robotics Inc. (https://reverbrobotics.ca). The single-board computer was equipped with custom audio hardware, comprised of four high-resolution digital microphones and speaker output. The microphones enabled ABRA to capture voice input for speech recognition, whereas the speakers played audio, and executed text-to-speech functions. To facilitate user interaction, the board, microphones, and speaker were housed within a custom 3D-printed case (see Fig. 2) so that users could engage with it as a self-contained unit.

2.2.2. Software

The provided system bundles state-of-the-art artificial neural

networks for speech recognition, natural language understanding, and speech output to allow two-way voice interaction between humans and machines (see whitepaper for details: https://reverbrobotics.ca/whitepaper/).

We implemented the memory tests using the Python programming language (https://github.com/reverbrobotics/seniors_rero_ai_experiment) and integrated with the board's provided API. We structured each memory test using a state machine. The process began with an instruction state followed by a confirmation state that would continue until the participant confirmed that they understood the instructions (see Fig. 1). After this confirmation, the experiment consisted of prompts from ABRA, each requiring a response from the user. ABRA played all prompts using its on-board Text-To-Speech functionality, and captured participant responses through its built-in speech recognition system (Vosk; Alpha Cephei Inc. (2023)). During the experiment, we used a feature of Vosk that limits the accepted vocabulary from the participant, depending on the task being completed. For example, responses were limited to digits during the backward and sequencing digit span tasks. Responses and timing information were stored in a text file using CSV format. Audio recordings of each response were also saved to ABRA for post-hoc analysis. All raw data (sans audio recordings for privacy) are available at https://osf.io/9m586/

2.3. Cognitive tasks

Three memory tasks were administered through the ABRA. ABRA was placed in the middle of a tabletop, located in front of the participant (see Fig. 2; its location did not change between participants). Three working memory verbatim tasks were administered; 1) logical memory task (WAIS-III-CDN; Wechsler, 2001), 2) poem recall task, 3) backward and sequencing digit span task (WAIS-IV-CDN; Wechsler., 2008). These specific neuropsychological tests were chosen for three main reasons: First, the Wechsler neuropsychological tests have been recognized as some of the most effective tools for classifying stages of none, mild, or severe cognitive impairment (Battista, Salvatore, & Castiglioni, 2017). Second, these are also among the recommended measures for cognitive assessment of older adults with dementia (Bossers, van der Woude, Boersma, Scherder, & van Heuvelen, 2012). Third, it was important to tap on the distinct types of verbal memory. The logical memory task primarily assesses both short-term and long-term verbal episodic memory, which involves the encoding, storage, and retrieval of verbal information that is associated with a particular event or experience. In contrast, the poem recall task focuses on short-term memory, particularly the ability to recall verbal information verbatim without significant processing or contextual manipulation. The backward and sequencing digit span task, on the other hand, measures verbal working memory, as it requires participants to not only hold verbal information temporarily but also actively manipulate and transform that information. This task differentiates itself from the others by engaging cognitive processes related to mental manipulation and sequencing, rather than simple storage or recall.

Fig. 2. Participant setup.
Note. Consent was obtained from the participant for use of this image for publication.

2.3.1. Logical memory

The task consisted of three sections: logical memory I, and logical memory II. All three parts are based on two stories (A and B). A delay time of 20–30 min is given between logical memory I and the other two sections. The logical memory I consists of an immediate recall of both stories. The logical memory II consists of a delayed (after 20–30 min) recall of the stories.

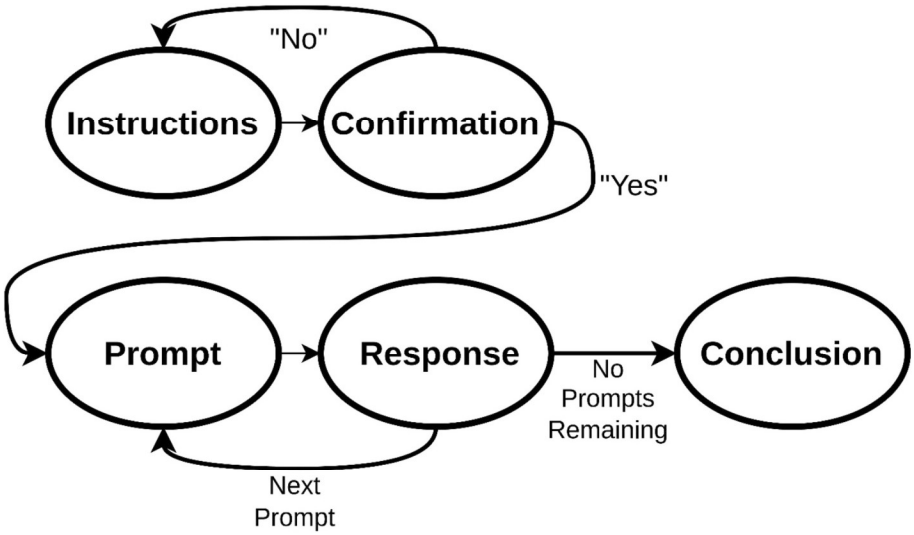


Fig. 1. Diagram of the basic state machine used throughout the experiment.
Note. Instructions are given to the participant, and once confirmed will move on to the Prompt. A set of a Prompts are given to the user, each requiring a Response. Once all Prompts have been exhausted, the experiment is concluded.

2.3.2. Poem recall

The task consisted of three different poems ("Invention" by Billy Collins, "The moment" by Margaret Atwood, and "Stopping by woods on a snowy evening" by Robert Frost). Each poem was broken down into sentences of similar length. ABRA communicated this poem in a sentence-by-sentence structure. The participant's job was to recall the sentence as accurately as possible and to recite it back.

2.3.3. Backward and sequencing digit span

The task consisted of two sections. The backward digit span and the sequencing digit span. The backward digit span consists of a series of digits (between 1 and 9) that need to be recalled by the participant and communicated verbally in reverse order. For example, if the series of digits where 1 and 2, the participant would answer 2 and 1. Per trial, the number of digits ranged from 2 and up to 8. Two trials were given per block, with the test ending if the participant got the two trials wrong. The sequencing digit span was similar to the backward digit span, except the digits given had to be repeated in order (starting from the lowest). For example, if the series of digits were 2–4-1; the participant would answer 1–2-4. In this case, per trial, the number of digits ranged from 2 and up to 9. The task also ended when the participant incorrectly answered two of the trials.

The experiment was broken down into four parts, administered in the same order for all participants: 1) logical memory task (part I), 2) poem recall task, 3) backward and sequencing digit span task, 4) logical memory task (part II).

2.4. Procedure

Participants were first asked to read and sign a consent form. As well, they were asked to fill out a demographic form which asked their personal details (contact information, sex/gender, race and ethnicity, marital status, and educational level). Participants were then seated at a table in front of ABRA (see Fig. 2). Once the participant was ready, the researcher prompted the AI device to start the experiment. Note that the experiment pace was based on the participant; ABRA continuously checked for participant readiness, if the participant was not ready to commence (or continue) ABRA would wait and ask again (the experiment would not continue until the participant was ready).

ABRA started the first part of the experiment with the set of instructions of the first section of the logical memory I-immediate recall (i.e., "I am going to read a short story to you. Listen carefully and try to remember it just the way I say it, as close to the same words as you can remember. When I am through, I want you to tell me everything I read to you. You should tell me all you can remember even if you are not sure"). Further, ABRA followed with questions to confirm understanding (i.e., "Do you understand?") of the instructions given previously and to check for participants readiness (i.e., "Let me know when you are ready to start"). Story A was then told by the AI device and the participants replied with as much as they remembered from the story. After, ABRA checked for any missing information (i.e., "Is there anything else you want to add?"), if there was any, the participant replied, and if not, ABRA continued onto story B, the same follow up question after story A, followed story B.

Further on, ABRA continued to the second part of the experiment, in which instructions of task 2 (poem recall) were given to participants (i.e., "I will read a poem, when I pause you may recite back what I said as accurately as possible. After you recite it back, we will continue to the next verse, until the poem is finished"). As in part one, ABRA checked for participant's understanding (i.e., "Do you understand?") and readiness (i.e., "Let me know when you are ready to start"). ABRA moved on in a sentence-by-sentence structure; for example, ABRA said: "Tonight the moon is a cracker" to what the participant was expected to respond: "Tonight the moon is a cracker", and so on, until the three different poems finished.

ABRA then followed into part three of the experiment, in which the set of instructions for task 3 section 1 (backward digit span) were given to participants (i.e., "I am going to say some numbers. Listen carefully, I can only say them one time. When I am through, I want you to say the numbers backward. For example, if I say 7, 1, then you would say "1, 7."). Once again, ABRA checked for participant's understanding and readiness. Once the participant was ready the task began. The task continued until the participant missed two of the trials with the same number of digits, at which point ABRA stopped the task. ABRA then followed onto section 2 (sequencing digit span) set of instructions (i.e., "Now I am going to say some more numbers. After I say them, I want you to tell the numbers in order, starting with the lowest number. If I say: 2, 4, 1. Then, you would say: 1, 2, 4"). Like in section 1, ABRA stopped the task when the participant missed two consecutive trials with the same number of digits.

In part four of the experiment, ABRA first began with the set of instructions of the logical memory II-delayed recall (i.e., "Do you remember the stories I read to you a little while ago? I want you to tell me the stories again. Tell me everything that you can remember about the first story. Start at the beginning"). ABRA checked if the participants needed a hint about the story A (i.e., Do you need a hint about the story topic?"). If participants needed a hint, then ABRA told them the hint (i.e., "The story was about a woman who was robbed") if they did not, the participant would answer "no" and then ABRA continued (i.e., "Now tell me everything that you can remember about the first story"). Participants answered everything they remembered about the first story (i.e., story A). After they were finished, ABRA checked for any missing information (i.e., "Is there anything else you want to add?"). For the second story (i.e., story B), the same procedure were followed as for story A, except, the hint was different (i.e., "The story was about a weather bulletin").

2.5. Data analysis

2.5.1. Logical memory

The logical memory immediate recall (part I) and logical memory delayed recall (part II) consisted of two scores (unit and thematic). The "unit" score accounted for the participants' ability to recall the details of the story, whereas the "thematic" score accounted for the participants' ability to recall the overall theme of the story. One point was awarded for each response. A total unit score of 25 points for each story (A and B) was awarded. Therefore, for part I-immediate recall a perfect score was 50 points and for part II-delayed recall a perfect score was of 50 points. A total thematic score of 15 points for part I-immediate recall were given and a total of 15 points were given for part II- delayed recall. Participants performance before and after were examined with paired-samples *t*-tests.

2.5.2. Poem recall

Number of verbatim errors were counted. Three types of errors were recorded: omission, substitution, and addition errors. An omission error was recorded when the participant missed one word during the sentence recall. For example, the sentence given was: "It will waste away to nothing, **nothing** but stars in the sky" and the participant answered: "It will waste away to nothing, but stars in the sky". A substitution error was awarded when the participant substituted a word for a similar word during the sentence recall. For example, the sentence given was: "His house is in the village **though**" and the participant answered: "His house is in the village **of**". An addition error was recorded when the participant added a word to the sentence. For example, the sentence given was: "knowing at last how you got there" and the participant answered: "knowing at last how you **know** you got there". A repeated measures ANOVA was used to investigate differences between errors. Significant main effects were analyzed with post-hoc comparisons, and the familywise error rate was controlled with the Bonferroni correction.

2.5.3. Backward and sequencing digit span

For both tasks, two scores were recorded: the "sum of the items" score and the "longest span" score. Participants received one point for each series correctly recalled and zero points if they made a mistake. The maximum "sum of the items" score was 16 because there were eight sequences for each, the backward and the sequencing digit span. The length of the maximum number of digits recalled made up the "longest span" score. For example, if the series given was "1–3–6-7-8-4", and the participant answered it correctly but made two consecutive mistakes on the next series (i.e., "1–3–6-7-8-4-9"), the participant would be awarded a score of 6, for the "longest span" score. The "longest span" given for the sequencing span was nine and for the backward span was eight. Participants performance in the sequencing and backward tasks were examined with paired-samples *t*-tests.

Because the tests in this study assessed primarily memory, and memory declines with age, we investigated if the results from any of the tests were correlated with age using a bivariate correlational analysis (Pearson's r). Age was included with the results of all the tests (unit score - immediate recall, thematic score-immediate recall, unit score-delayed recall, thematic score- delayed recall, sum of Items score sequencing, longest span sequencing, sum of items score backward, longest span backward, omission errors, substitution errors, addition errors) as dependent variables. Alpha level for all comparisons was 0.05.

3. Results

3.1. Cognitive tasks

Sex was accounted for as a variable of interest in the study, however, there were not main effects or interactions with any of the dependent variables. Therefore, this independent variable was excluded from all analyses.

3.1.1. Logical memory test

Grand mean scores are shown in Table 1. Paired sample *t*-tests were conducted, to compare the two scores (immediate and delayed recall) from the unit and thematic scores of the logical memory task. There was a significant difference between the immediate recall and the delayed recall concerning the unit scores and an approaching significant difference in the thematic scores (see Table 1). This suggested that participants recalled better the details and theme of the story immediately after it was read to them when compared to 20–30 min after.

3.1.2. Poem recall

Grand means for all omission, substitution, and errors are shown in Table 2. A repeated measures ANOVA was used to investigate differences among the different types of errors, and the familywise error rate was controlled with the Bonferroni correction. The analysis was significant ($F_{(2,46)} = 58.38$, $p < 0.001$, $\eta^2 = 0.7$). Participants made significantly more omission errors compared to substitution errors, and participants made more substitution errors when compared to addition errors (all comparisons were significantly different from each other $p < 0.001$).

3.1.3. Backward and sequencing digit span

Grand mean scores are shown in Table 3. Paired sample t-tests were conducted to compare differences between the sequencing and backward digit tasks, concerning the sum of the items and the longest span scores. There was a significant difference between the sequencing and backward digit tasks. As hypothesized, participants remembered more digits and achieved a longer span in the sequencing compared to the backward digit task.

Table 1
Means and standard errors for the dependent variables.

Dependent variables	Immediate recall	Delayed recall	*t* statistics
Unit score	22.9 ± 2.0	18.5 ± 1.5	$t_{(23)} = 6.1$, $p < 0.001$, $d_s = 1.24$
Thematic score	10.9 ± 0.5	10.1 ± 0.6	$t_{(23)=} 1.9$, $p = 0.06$, $d_s = 0.4$

Table 2
Means and standard errors for the dependent variables.

Dependent variables	Values
Omission errors	34 ± 3.9
Substitution errors	8.9 ± 0.9
Addition errors	2.3 ± 0.5

Table 3
Means and standard errors for the dependent variables.

Dependent variables	Backward digit	Sequencing digit	*t* statistics
Sum of items score	7.4 ± 0.28	6.4 ± 0.31	$t_{(23)} = 3.03$, $p < 0.01$, $d_s = 0.62$
Longest span score	3.9 ± 0.2	4.5 ± 0.18	$t_{(23)=} - 3.02$, $p < 0.01$, $d_s = -0.62$

3.2. Correlation analysis

Table 4. shows the results of the correlation analysis. There were significant negative correlations (all $ps < 0.05$) between age and performance in the unit score-delayed recall, thematic score-delayed recall, sum of items score- sequencing, and the longest span score-sequencing; the older the participant the lower their score or span (worse performance). Correlations between age and unit score-immediate recall and thematic score-immediate recall were not significant, although all $ps < 0.1$ were in the right direction (with an increase in age a worse performance at the task). No significant correlations were found between age and the sum of items score-backward, longest span score-backward, omission errors, substitution errors, or addition errors.

3.3. How does the AI compare to human administration of the logical memory test?

Data ($N = 24$) from a previous study with a human administrating the Wechsler logical memory test (Coelho et al., 2020) allowed for a comparison with the administration of this test using the ABRA (current study). A semi-random (matched for age) cohort of participants were chosen from Coelho et al.' (2020) dataset. A participant sample between the ages of 60 to 85 (M = 702, SD = 5.8, Mdn = 70.5) were included for comparison purposes. Independent sample *t*-tests revealed no differences in age between the human and the ABRA groups ($t_{(46)=} - 0.9$, $p = 0.4$, $d_s = -0.3$). Logical memory test immediate and delayed recall scores were compared between the samples matched by number of participants. Group differences were investigated with independent samples t-tests for each score. There were no significant differences between the human and ABRA administration in unit and thematic scores for both immediate or delayed recall-(all $ps > 0.2$).

4. Discussion

As the global population ages, researchers and health care professionals must seek strategies to lessen the strain on health care resources without compromising quality of care. Maintaining and preferably enhancing the standard of care for older adults amidst this crisis is crucial. AI systems are essential tools that can help tackle current and upcoming health care challenges. AI has proven useful to societies worldwide, by achieving complex tasks with unparalleled precision and

Table 4
Correlation matrix for the dependent variables.

	Wechsler Logical Memory Test				Wechsler Digit Span Test				Poem Recall Task		
	US-IR	TS-IR	US-DR	TS-DR	SI-B	LS-B	SI-S	LS-S	OE	SE	AE
Age	−0.37*	−0.36*	−0.47**	−0.53***	-0.63***	−0.57***	−0.28	−0.26	0.28	0.11	0.32

Note. US-IR (Unit Score- Immediate Recall), TS-IR (Thematic Score- Immediate Recall), US-DR (Unit Score- Delayed Recall), TS (Thematic Score- Delayed Recall), SI-B (Sum of Items-Backward), LS-B (Longest Span-Backward), SI-S (Sum of Items-Sequencing), LS-S (Longest Span-Sequencing), OE (Omission Errors), SE (Substitution Errors), AE (Addition Errors). $p < 0.1$*, $p < 0.05$*, $p < 0.01$**.

speed. Unfortunately, not much work has been dedicated to developing AI systems that acknowledge an older population by addressing their specific needs, and their sensory and cognitive capabilities. Therefore, in the current study, we aimed to administer well-established neuropsychological tests via an AI-powered speech device that allowed for a two-way voice communication between the older adult and the system. We hypothesized that ABRA would reliably collect data from an older population enabling the assessments of these tests without human intervention. Introducing such a device in research and healthcare settings would offer three advantages. First, AI devices can facilitate more efficient (i.e., fast and precise) administration of cognitive tests than human-led collection. Second, AI devices can increase productivity by enabling researchers and healthcare professionals to allocate more attention to tasks requiring human guidance and interaction. Third, the possibility of remote assessments becomes possible through the use of AI devices, alleviating access concerns among older adults living in rural and remote areas.

In the current study, three memory tasks were administrated via the ABRA: the logical memory task (WAIS-III-CDN; Wechsler, 2001), the poem recall task, and the digit span task (WAIS-IV-CDN; Wechsler., 2008). The logical memory and the digit span tests are two well-established and commonly used neuropsychological tests which have proven sensitive at distinguishing between healthy aging, compared to mild cognitive impairment, and those with more severe cognitive impairment like Alzheimer's (Bruno et al., 2021; Chapman, Anand, Sparks, & Cullum, 2006; Kessels, Overbeek, & Bouman, 2015) and Parkinson's disease (Cooper, Sagar, Jordan, Harvey, & Sullivan, 1991; Werheid et al. 2002, Warden, Hwang, Marshall, Fenesy, & Poston, 2016). Despite their established effectiveness, to our knowledge, these tests have not been previously administered using an automatic two-way voice communication speech technology. Our hypothesis was supported, the ABRA proved reliable for administering all tests used in this study. The semantic verbal fluency test has been mostly administered using automatic speech recognition and speech-to-text in young adults (Pakhomov, Marino, Banks, & Bernick, 2015) and older adults (Konig et al., 2018; Troger et al. 2019). Taken together these studies and the results from the present investigation have produced positive results, demonstrating automatic speech recognition is feasible for the assessment of neuropsychological function.

The results of the logical memory test showed that older adult participants had better recall regarding the details of the story (unit score) immediately after ABRA's presentation when compared to the delayed recall of the stories (after 20–30 min). These results are in line with human administration of this test (Bruno et al., 2021). Previous research performed in a large database of participants, has shown that there is a significant main effect between the two time points of the unit score (immediate versus delayed recall) on participants that are cognitively healthy, with cognitive decline, or mild cognitive impairment (Bruno et al., 2021). Regarding the thematic score, a significant difference between the two time points was not found on the current study. This result has been found before with human administration of the test. For example, Chapman et al. (2006), found that healthy "young-old" and "old-old" adults recalled the overall theme of the stories, with minimal change in performance from the immediate and delayed thematic scores. Since the major work of British psychologist Bartlett (1932), it was suggested that individuals remember the main idea of the story better than the details even at different delayed time points. Other researchers have more recently supported these findings, where older adults recall better the gist of the story rather than the details (Adams, Smith, Nyquist, & Perlmutter, 1997; Barber & Mather, 2014). It appears that the results of the thematic portion of the test, are similar when administered through ABRA; participants recalled as much information immediately after or with a 20–30 min delay.

When asked to recall verbatim stories or (in this case) more specifically, poems, older adult participants omitted words (omission errors), substituted words (substitution errors), and added words (addition errors). They significantly omitted more words than they substituted, and they significantly substituted more words than they added. Davis, Alea, and Bluck (2015) observed similar results in an older adult group, when asked to recall stories, participants made more errors of omission than errors of substitution or addition (or as they called them change errors and errors of commission, respectively). Interestingly, Davis et al. (2015), also showed that details that were perceptual (e.g., "fireworks were spectacular") or emotional (e.g., "impressed with the evening") in nature, tended to be omitted more frequently in the older adult population. Poems tend to contain more perceptual and emotional details; thus, this may be one of the explanations behind a greater number of omission errors.

The results from the digit span tests assessed via ABRA, showed that participants recalled more items in the backward digit span task when compared to the sequencing digit span, but participants achieved a longer span in the sequencing digit span when compared to the backward digit span. A research study in a large sample ($N = 1205$) of older adults revealed similar results (LaBelle, Lee, & Miller, 2019). Older adults recalled significantly more items (or total number of sequences) in the backward digit span compared to the sequencing digit span (M = 7.7 versus M = 6.20, respectively) but the span was significantly longer for the sequencing digit span when compared to the backward digit span (M = 4.64 versus M = 4.13). These researchers suggest that the longest span seen on sequencing digit span may be due to having an already established knowledge of the natural number sequence (e.g., 1, 2,3, 4…). It is important to note that the mean scores of the backward digit span (M = 7.5) and SDS (M = 6.7) administered through ABRA, are in line with the mean scores from this and other studies in older adults (Werheid et al., 2002; Wisdom, Mignogna, & Collins, 2012).

Furthermore, our results of the correlation analysis align with previous research showing that age is negatively correlated with performance on the Logical Memory and Digit Span tests of the WAIS-III-CDN and WAIS-IV-CDN batteries, respectively. The scores on the delayed portions of the logical memory test were negatively associated with participants' age. In two large ($N = 1250$) census studies with participants between the ages of 16 and 89, researchers found an age associated deterioration in immediate and delayed recall of this test (Haaland, Price, & Larue, 2003; Price, Said, & Haaland, 2004). Furthermore, age was also significantly negatively correlated with the sum of items and longest span for the sequencing digits task. Overall, the digits task is suggested to be an important measure in neuropsychological testing in patients with dementia (Bossers et al., 2012) and has been associated with diagnosis of neurocognitive disorder in a sample of veterans (Lumpkin & Sheerin, 2019). Thus, on these two tests of the Wechsler battery, ABRA was a sensitive tool to identify age-related changes even with a relatively small sample. Notably too, is the finding that the score

of the logical memory test (both immediate and delayed) when collected by ABRA agree with those administered by a human experimenter (Coelho et al., 2020). Together, this further supports our premise that the ABRA can perform analogous to a human examiner.

The implementation of ABRA in the neuropsychological assessment of older adults presents several challenges that warrant careful consideration. A central issue stems from the current generation of older adults' limited knowledge and understanding of AI and its mechanisms. Giannouli (2023) argues that this knowledge gap may be contributing to feelings of apprehension and distrust, with many participants expressing a clear preference for human examiners over AI-administered assessments (Giannouli, 2023). With respect to our study, some participants expressed similar concerns "human interaction is preferred", but others indicated that they would use the device for cognitive games, medication tracking, and daily health checks. Nevertheless, if ABRA was to be integrated into healthcare settings, these concerns must be addressed to ensure effective and ethical application. Encouragingly, there is evidence that older adults demonstrate a willingness to learn about AI and its practical applications (Giannouli, 2023). Therefore, it is crucial to develop targeted educational programs that demystify AI, specifically explaining how devices such as ABRA function and contribute to assessments may reduce negative perceptions and promote greater acceptance. Fostering this understanding will be critical to ensuring the successful adoption of AI-based tools in neuropsychological practice.

Employment of automatic speech recognition devices for cognitive assessment in older adults yields numerous benefits for research and health care areas, beyond mere efficiency gains for health care professionals. First, assessment via AI offers a natural and comparable way to an in-person test administrator. Second, use of a speech device breaches any unfamiliarity with technological devices the older adult patient or participant may have (Sirilertmekasakul et al., 2023). Third, and most importantly, remote administration enables inclusion of hard-to-reach populations and facilitates the continuous screening of older adults. This in turn, could help early identification of prodromal stages of neurodegenerative diseases as Alzheimer's disease and Parkinson's disease. Finally, ABRA's dynamic, real-time response capabilities open the door to implementing tasks that are more ecologically valid, not only in verbal assessments but also in other domains such as spatial and executive functions (Tragantzopoulou & Giannouli, 2024).

In sum, the current study demonstrated that the assessment of neuropsychological tests via ABRA an AI-powered speech device is reliable, and analogous to the standard human-administered version in an older population. While this study was investigatory in nature, we were pleased to find that the barriers of use were lower than we expected; the older population was capable of understanding and using the technology. Future studies should focus on examining the reliability of other neuropsychological tests administered through this device or similar speech devices. Harnessing the potential of this technology can assist us in addressing present and future health care challenges more effectively.

4.1. Limitations

A notable limitation of the current investigation was that all the tests were manually scored by a human. Neuropsychological evaluations have and continue to be primarily executed by trained human examiners. Although care was taken to ensure consistency and accuracy in scoring the current results, manual scoring introduces the possibility of human error and subjective bias. This method is also more time-consuming and resource-intensive compared to automated methods. Potential future versions of ABRA should be end-to-end in that collection and scoring is done automatically. We are currently incorporating the necessary tools to ABRA, so it can accurately analyze the data. We believe researchers and health care professionals will greatly benefit from this, as it will drastically reduce workload requirements.

CRediT authorship contribution statement

Daniela E. Aguilar Ramirez: Formal analysis, Data curation, Writing – original draft. **Lukas Grasse:** Software, Methodology, Writing – review & editing. **Scott Stone:** Software, Writing – review & editing. **Matthew Tata:** Supervision, Methodology. **Claudia L.R. Gonzalez:** Supervision, Methodology, Writing – review & editing.

Funding

This project was funded by the Natural Sciences and Engineering Research Council of Canada with a Tier II Canada Research Chair and a Discovery Grant awarded to Dr. Claudia Gonzalez (Grant no. 14367).

Declaration of competing interest

The authors (LG, SS, and MT) disclose an affiliation with Reverb Robotics Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors thank all participants for their contribution to this project.

The authors would like to thank the funding agency for supporting this research.

Data availability

Data is available at a repository (link provided on the manuscript and on the data availability statement)

References

Adams, C., Smith, M. C., Nyquist, L., & Perlmutter, M. (1997). Adult age-group differences in recall for the literal and interpretive meanings of narrative text. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 52*(4), P187–P195. https://doi.org/10.1093/geronb/52b.4.p187

Alpha Cephei Inc.. (2023). Vosk offline speech recognition API. https://alphacephei.com/vosk/.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., … Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine, 25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Barber, S. J., & Mather, M. (2014). How retellings shape younger and older adults' memories. *Journal of Cognitive Psychology (Hove, England), 26*(3), 263–279. https://doi.org/10.1080/20445911.2014.892494

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge University Press.

Battista, P., Salvatore, C., & Castiglioni, I. (2017). Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study. *Behavioural Neurology, 2017*, 1850909. https://doi.org/10.1155/2017/1850909

Bossers, W. J. R., van der Woude, L. H. V., Boersma, F., Scherder, E. J. A., & van Heuvelen, M. J. G. (2012). Recommended measures for the assessment of cognitive and physical performance in older patients with dementia: A systematic review. *Dementia and Geriatric Cognitive Disorders Extra, 2*(1), 589–609. https://doi.org/10.1159/000345038

Bruno, D., Mueller, K. D., Betthauser, T., Chin, N., Engelman, C. D., Christian, B., … Johnson, S. C. (2021). Serial position effects in the logical memory test: Loss of primacy predicts amyloid positivity. *Journal of Neuropsychology, 15*(3), 448–461. https://doi.org/10.1111/jnp.12235

Chapman, S. B., Anand, R., Sparks, G., & Cullum, C. M. (2006). Gist distinctions in healthy cognitive aging versus mild alzheimer's disease. *Brain Impairment, 3*, 223–233.

Chen, S., Stromer, D., Alabdalrahim, H. A., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports, 10*(1), 20854. https://doi.org/10.1038/s41598-020-74710-9

Coelho, L., Hauck, K., McKenzie, K., Copeland, J. L., Kan, I. P., Gibb, R. L., & Gonzalez, C. L. R. (2020). The association between sedentary behavior and cognitive ability in older adults. *Aging Clinical and Experimental Research, 32*(11), 2339–2347. https://doi.org/10.1007/s40520-019-01460-8

Cooper, J. A., Sagar, H. J., Jordan, N., Harvey, N. S., & Sullivan, E. V. (1991). Cognitive impairment in early, untreated Parkinson's disease and its relationship to motor disability. *Brain, 114*(Pt 5), 2095–2122. https://doi.org/10.1093/brain/114.5.2095

Davis, D. K., Alea, N., & Bluck, S. (2015). The difference between right and wrong: Accuracy of older and younger adults' story recall. *International Journal of Environmental Research and Public Health, 12*(9), 10861–10885. https://doi.org/10.3390/ijerph120910861

Dean, J. (2022). A golden decade of deep learning: Computing systems& applications. *Daedalus, 151*(2), 58–74. https://doi.org/10.1162/daed_a_01900

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Giannouli, V. (2023). Financial capacity assessments and AI: A Greek drama for geriatric psychiatry? *International Journal of Geriatric Psychiatry, 38*(9), Article e6008. https://doi.org/10.1002/gps.6008

Glyde, H., Hickson, L., Cameron, S., & Dillon, H. (2011). Problems hearing in noise in older adults: A review of spatial processing disorder. *Trends in Amplification, 15*(3), 116–126. https://doi.org/10.1177/1084713811424885

Guo, J., Huang, X., Dou, L., Yan, M., Shen, T., Tang, W., & Li, J. (2022). Aging and aging-related diseases: From molecular mechanisms to interventions and treatments. *Signal Transduction and Targeted Therapy, 7*(1), 391. https://doi.org/10.1038/s41392-022-01251-0

Haaland, K., Price, L., & Larue, A. (2003). What does the WMS-III tell us about memory changes with normal aging? *Journal of the International Neuropsychological Society : JINS, 9*, 89–96. https://doi.org/10.1017/S1355617703910101

Kaiser, A. K., Hitzl, W., & Iglseder, B. (2014). Three-question dementia screening. Development of the Salzburg dementia test prediction (SDTP). *Zeitschrift für Gerontologie und Geriatrie, 47*(7), 577–582. https://doi.org/10.1007/s00391-013-0568-7

Kessels, R. P. C., Overbeek, A., & Bouman, Z. (2015). Assessment of verbal and visuospatial working memory in mild cognitive impairment and alzheimer's dementia. *Dement Neuropsychol, 9*(3), 301–305. https://doi.org/10.1590/1980-57642015DN93000014

Kim, Y. H. (2003). *Geriatric speech. Plenary session IV. Yonsei University College of Medicine, otolaryngology clinic, 205–207.*.

König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., & Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and Geriatric Cognitive Disorders, 45*(3–4), 198–209. https://doi.org/10.1159/000487852

Kwon, S., Kim, S. J., & Choeh, J. Y. (2016). Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language, 36*, 110–121. https://doi.org/10.1016/j.csl.2015.09.002

LaBelle, D. R., Lee, B. G., & Miller, J. B. (2019). Dissociation of executive and attentional elements of the digit span task in a population of older adults: A latent class analysis. *Assessment, 26*(7), 1386–1398. https://doi.org/10.1177/1073191117714556

Lee, S. Y. (2011). *The overall speaking rate and articulation rate of normal elderly people. Graduate program in speech and language pathology, master these, Yonsei University.*.

Lumpkin, J. C. M., & Sheerin, C. M. (2019). Digit span sequencing as a neurocognitive screening tool in an aging, veteran population. *Psychology & Neuroscience, 12*(2), 180–190. https://doi.org/10.1037/pne0000140

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *ArXiv*. abs/1312.5602.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Pacala, J. T., & Yueh, B. (2012). Hearing deficits in the older patient: "I didn't notice anything". *JAMA, 307*(11), 1185–1194. https://doi.org/10.1001/jama.2012.305

Pakhomov, S. V. S., Marino, S. E., Banks, S., & Bernick, C. (2015). Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency. *Speech Communication, 75*, 14–26. https://doi.org/10.1016/j.specom.2015.09.010

Park, I., Kim, Y. J., Kim, Y. J., & Lee, U. (2020). Automatic, qualitative scoring of the interlocking pentagon drawing test (PDT) based on u-net and mobile sensor data. *Sensors (Basel), 20*(5). https://doi.org/10.3390/s20051283

Price, L., Said, K., & Haaland, K. Y. (2004). Age-associated memory impairment of logical memory and visual reproduction. *Journal of Clinical and Experimental Neuropsychology, 26*(4), 531–538. https://doi.org/10.1080/13803390490496678

Sirilertmekasakul, C., Rattanawong, W., Gongvatana, A., & Srikiatkhachorn, A. (2023). The current state of artificial intelligence-augmented digitized neurocognitive screening test. *Frontiers in Human Neuroscience, 17*, 1133632. https://doi.org/10.3389/fnhum.2023.1133632

Tragantzopoulou, P., & Giannouli, V. (2024). Spatial orientation assessment in the elderly: A comprehensive review of current tests. *Brain Sciences, 14*(9), 898. https://doi.org/10.3390/brainsci14090898

Trajkova, M., & Martin-Hammond, A. (2020). *"Alexa is a toy": Exploring older adults' reasons for using, limiting, and abandoning echo. Proceedings of the 2020 CHI conference on human factors in computing systems, 1–13*. https://doi.org/10.1145/3313831.3376760

Tröger, J., Linz, N., König, A., Robert, P., Alexandersson, J., Peter, J., & Kray, J. (2019). Exploitation vs. exploration-computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia, 131*, 53–61. https://doi.org/10.1016/j.neuropsychologia.2019.05.007

United States Census Bureau. (2018, March). *13*. History: Older People Projected to Outnumber Children for First Time in U.S. https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html.

Völter, C., Thomas, J. P., Maetzler, W., Guthoff, R., Grunwald, M., & Hummel, T. (2021). Sensory dysfunction in old age. *Deutsches Ärzteblatt International, 118*(29–30), 512–520. https://doi.org/10.3238/arztebl.m2021.0212

Warden, C., Hwang, J., Marshall, A., Fenesy, M., & Poston, K. L. (2016). The effects of dopamine on digit span in parkinson's disease. *J Clin Mov Disord, 3*, 5. https://doi.org/10.1186/s40734-016-0033-z

Wechsler, D. (2001). *Wechsler Adult Intelligence Scale- Third edition: Canadian technical manual*. Toronto, ON: Harcourt Canada.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition: Canadian Technical Manual*. Toronto, Ontario: Pearson Canada.

Werheid, K., Hoppe, C., Thöne, A., Müller, U., Müngersdorf, M., & von Cramon, D. Y. (2002). The adaptive digit ordering test clinical application, reliability, and validity of a verbal working memory test. *Archives of Clinical Neuropsychology, 17*(6), 547–565. https://doi.org/10.1093/arclin/17.6.547

Wisdom, N. M., Mignogna, J., & Collins, R. L. (2012). Variability in Wechsler adult intelligence scale-IV subtest performance across age. *Archives of Clinical Neuropsychology, 27*(4), 389–397. https://doi.org/10.1093/arclin/acs041

World Health Organization. (2022, October). 1. Ageing and health https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

Yu, G., Sun, K., Xu, C., Shi, X.-H., Wu, C., Xie, T., Meng, R.-Q., Meng, X.-H., Wang, K.-S., Xiao, H.-M., & Deng, H.-W. (2021). Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. *Nature Communications, 12*(1), 6311. https://doi.org/10.1038/s41467-021-26643-8