



A preliminary psychometric evaluation of the activity ordering task with a metacognitive facet (AOT-M)

Nidhi Lalu Jacob¹ Aysha Rooha¹ Anjaly S. Nair² Gagan Bajaj¹ Vinitha Mary George³ Jayashree S. Bhat⁴ 

Keywords

Aging
Metacognition
Working Memory
Reliability
Validity

ABSTRACT

Purpose: The Activity Ordering Task with a metacognitive facet (AOT-M) was developed, in our previous work, to address the disconnect between traditional working memory (WM) tasks and everyday WM demands, the lack of culturally sensitive, context-based WM tasks in India and enhance participant engagement. The present study aims to provide preliminary evidence of the AOT-M's psychometric properties among a non-clinical adult population, evaluate its sensitivity to cognitive and metacognitive changes with aging, establish construct validity, ecological validity, concurrent validity and test-retest reliability. **Methods:** Ninety neurotypical adults, evenly distributed across three age groups, participated in the study. Descriptive statistics examined the distribution of performance spans and estimation discrepancies across age groups and the age-related statistical differences were evaluated using the Kruskal-Wallis Test. Construct validity was assessed using Rasch analysis, while ecological validity was evaluated with the Multidimensional Assessment of Research in Context (MARC) tool. Concurrent validity with sentence ordering and digit letter ordering tasks, was determined through Pearson's correlation coefficient and test-retest reliability was assessed using the Intraclass Correlation Coefficient and Bland-Altman plots. **Results:** The patterns observed in WM performance spans and estimation discrepancies highlighted the task's sensitivity to aging related cognitive and metacognitive changes. Evidence from the MARC tool substantiated ecological validity, and concurrent validity was demonstrated through significant correlations with established WM tasks. While Rasch analysis supported construct validity, moderate person reliability indicated some limitations in task sensitivity. The AOT-M demonstrated good test-retest reliability. **Conclusion:** Overall, the study provides preliminary evidence of the AOT-M's good psychometric properties within a neurotypical adult sample, suggesting it to be a promising addition to the cognitive communicative toolbox for Speech Language Pathologists.

Correspondence address:

Gagan Bajaj
Department of Audiology and Speech
Language Pathology, Kasturba Medical
College Mangalore, Manipal Academy
of Higher Education
Manipal (Karnataka), India, 576104.
E-mail: gagan.bajaj@manipal.edu

Received: July 23, 2024

Accepted: October 23, 2024

Study conducted at Kasturba Medical College – KMC - Mangalore, Karnataka, India.

¹Department of Audiology and Speech Language Pathology, Kasturba Medical College Mangalore, Manipal Academy of Higher Education - Karnataka, Manipal, India.

²Division of Biostatistics, Malankara Orthodox Syrian Church Medical College & Hospital Kolenchery - Ernakulam, Kerala, India.

³Department of Audiology and Speech Language Pathology, National Institute of Speech & Hearing - Trivandrum, Kerala, India.

⁴Department of Audiology and Speech Language Pathology, Nitte Institute of Speech and Hearing, Nitte deemed to be University Deralakatte - Mangalore, Karnataka, India.

Financial support: The study is supported by funding received from the Department of Science and Technology (DST) under the Cognitive Science Research Initiative (CSRI), Government of India (DST/CSRI/2018/29).

Conflict of interests: nothing to declare.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

Working memory and metacognition in everyday life

In everyday life, working memory (WM) is essential for real-time information processing, language comprehension, the retention and manipulation of information necessary for executing complex tasks such as learning, reasoning, decision-making and effective social interactions⁽¹⁾. It plays a crucial role in facilitating a wide range of cognitive-communicative activities necessary for navigating daily routines across various phases of adulthood. For instance, in young adults (YA), WM is necessary for managing complex tasks such as following lectures, studying for exams, problem-solving during group projects, participating in workplace conversations and acquiring new skills⁽²⁾. Among middle-aged adults (MAA), WM is required for handling projects, organizing schedules, meeting deadlines and balancing work with personal obligations. WM facilitates older adults (OA) in managing daily activities such as medication, financial planning, meal preparation and social interactions⁽³⁾. WM deficits are prevalent across various cognitive-communicative disorders, including dementia, Mild cognitive impairment, aphasia and Traumatic brain injury, significantly impairing essential linguistic functions such as figurative language use, context integration in conversations, narrative coherence, pronoun resolution, comprehension and overall language processing⁽⁴⁾. Therefore, understanding and assessing WM across adulthood is fundamental for promoting cognitive resilience, building cognitive reserves and ultimately supporting healthy cognitive aging⁽⁵⁾.

Addressing the significance of WM across adulthood, numerous assessments have been developed and validated; however, traditional lab-based cognitive tasks such as Reading Span, Listening Span, Operation Span, Rotation Span, Digit Span and others^(6,7) often fail to capture the multifaceted WM demands of everyday life, leading to a lab-life gap⁽⁸⁾. Consequently, ecologically valid WM assessments like 'Shopping Mall Task' and 'Overnight Trip Task' have been developed, aiming to better predict functional performance in everyday contexts^(9,10). However, these ecologically valid tasks often lack culturally and linguistically relevant stimuli, frequently demand high-end technology and substantial time investment, which restricts their widespread adoption and usability⁽¹¹⁾. Given these considerations, there remains a notable scarcity of ecologically valid WM assessment measures tailored specifically for evaluating WM in Indian adults.

Another crucial aspect of cognitive measures in general, and WM specifically, is their potential to assess the metacognitive processes associated with them⁽¹²⁾. Specifically, metacognitive processes linked with WM significantly influence individuals' awareness of their cognitive abilities, impacting their performance in everyday communication scenarios⁽¹³⁾. Metacognition, described as 'thinking about thinking,' is essential across adulthood influencing decision-making, problem-solving, learning, error monitoring, strategy selection and social interactions⁽¹⁴⁾. It supports active learning, critical thinking, reflective judgment and efficient cognitive offloading, crucial for understanding one's

cognitive performance⁽¹⁵⁾. Assessing metacognition is vital in clinical and research settings to gauge how individuals perceive and manage their cognitive health, providing insights into their self-awareness of cognitive changes, aiding early detection and interventions to preserve cognitive function and quality of life throughout adulthood. Several laboratory-based cognitive tasks, including visuospatial WM assessments, listening span tasks, n-back tasks, reading span tasks, digit ordering tasks and operation span tasks are increasingly integrating metacognitive components^(12,16,17). These WM assessments utilize both offline measures, such as questionnaires and self-report tools, to explore global metacognition and online measures, such as prospective, concurrent, and retrospective measures, to provide dynamic metacognitive insights across cognitive tasks⁽¹⁸⁾. Despite these advancements, there is a dearth of ecologically valid WM tasks that embed comprehensive metacognitive components, specifically tailored to the Indian context.

Activity ordering task with a metacognitive facet (AOT-M)

Recognizing the need to address the disconnect between traditional WM task performance and everyday WM demands, participant engagement issues, the benefits of incorporating a metacognitive component and the challenges posed by existing context-based tasks that lack ease of utility and cultural sensitivity in the Indian context, the Activity Ordering Task with a metacognitive facet (AOT-M) was developed⁽¹⁹⁾. The AOT-M was designed following the Analysis, Design, Develop, Implementation and Evaluation (ADDIE) instructional design model, which provided a structured framework across five well established phases⁽²⁰⁾. The initial Analysis phase involved a thorough literature review to identify gaps and research questions. During the Design phase, the task was conceptualized using the Nominal Group Technique and underwent content validation. In the Develop phase, the content-validated script was computerized in collaboration with an animation artist and integrated into SuperLab software. Pilot testing during the Implementation phase further refined its usability. Our previous work details these first four phases of the task development comprehensively⁽¹⁹⁾.

The AOT-M is a progressive span-based assessment ranging from Level 2 to Level 10, featuring two trials per level to allow an additional attempt upon failure. Structured around everyday scenarios, participants must order activities chronologically based on instructions from various sources. This task requires participants to apply an overlearned ordering principle, actively maintaining both activities and timelines in primary memory until completion. As levels progress, the increasing cognitive demands may exceed the capacity of primary memory, requiring information to be stored in secondary memory for subsequent controlled retrieval. Participants use relevant cues, such as activity timelines, to recall details amidst distractions, continuously monitoring and updating WM representations to maintain accuracy. Attentional control processes are engaged to direct attention to task-relevant information and periodically refresh WM contents, ensuring the sequence of activities remains accurate and updated⁽²¹⁾. The AOT-M thus captures the intricate interplay of WM processes such as active maintenance, controlled

retrieval, monitoring, updating and inhibition. Additionally, the AOT-M incorporates a metacognitive facet where participants predict their WM span before and after completing the task, providing insights into their self-awareness of their performance. In view of the fact that WM facilitates language comprehension, expression and overall communication⁽⁴⁾, the AOT-M, which evaluates both WM and associated metacognition, could serve as a valuable tool for speech-language pathologists in cognitive-communicative assessments.

The present study

Establishing psychometric properties such as reliability and validity is fundamental in developing new measures for clinical practice, education and research, as they ensure confidence in the accuracy and interpretation of assessments^(22,23). Reliability ensures consistency and reproducibility across successive administrations whereas validity determines how well an instrument measures the intended construct⁽²⁴⁾. Typically, this validation process begins with non-clinical samples to establish initial utility before advancing to validation in clinical populations⁽²⁵⁾.

The evaluation phase of developing the AOT-M, following the ADDIE instructional design model, is designed as a comprehensive series of investigations beginning with non-clinical populations and progressing to clinical populations to assess the psychometric properties of the novel task. This study represents the first investigation in the series, focusing on providing preliminary evidence of the AOT-M's psychometric properties among non-clinical adult population. Specifically, the aims were to evaluate the AOT-M's sensitivity to cognitive and metacognitive changes associated with aging, assess construct validity using Rasch analysis, ascertain ecological validity, establish concurrent validity and evaluate its test-retest reliability.

METHODS

The present study focuses on the evaluation phase, presenting the initial psychometric properties of the AOT-M on a small, non-clinical adult population. Approval was obtained from the Institutional Ethics Committee (IECKMCMLR-08/2021/263, IECKMCMLR-05/2023/269).

Participants

In present research, 90 neurotypical adults were recruited through convenience sampling, with equal representation across

the three age groups: YA (18-40 years; male=7, female=23), MAA (41-65 years; male=8, female=22) and OA (> 65 years; male=13, female=17). Adults with a score of more than 26 on the Mini-Mental State Examination⁽²⁶⁾ having no history of neurological or psychological disorders were included as participants for this phase. The socioeconomic status of all participants was determined as middle class using the Modified Kuppuswamy scale⁽²⁷⁾. Participant's English proficiency was verified by ensuring a minimum proficiency score of 'seven' on the Language Experience and Proficiency Questionnaire⁽²⁸⁾. All participants signed and provided informed consent. Detailed demographic information for all participant groups is provided in Table 1.

Measures

AOT-M

AOT-M is a span-based WM measure comprising of two components aimed at assessing everyday WM and a metacognitive facet⁽¹⁹⁾. The task, presented in an audiovisual format, involves relatable everyday themes and requires participants to order activities for a character to be completed at various times of the day, based on instructions from family, friends, or colleagues. The primary objective is to arrange these activities in chronological order, from earliest to latest, upon receiving the prompt. The activities range from two to ten, with complexity levels equivalent to the participant's WM span. Each level includes two trials, providing participants a second chance if the initial attempt is unsuccessful. The WM span is recorded as the highest level at which participants can accurately order the activities in chronological order of their timelines. For instance, if a participant successfully orders 4 activities correctly but fails on two trials of 5 activities, their WM span is recorded as 4. An example of a trial is provided in Figure 1, displaying the task components. The metacognitive facet of the AOT-M employs an online method of assessing metacognition, where participants predict their WM span before performing the task and then postdict the highest WM span they believe achieved after completing the task. The metacognitive facet is scored by calculating the estimation discrepancy, which is the difference between the predicted or postdicted spans and the actual task performance span. The AOT-M is administered using SuperLab software, which supports simultaneous auditory-visual presentation with a 1000 millisecond inter-stimulus interval (ISI). The content validation of the AOT-M demonstrated high understandability scores of 90.9% for the script and 89.6% for the task⁽¹⁹⁾.

Table 1. Participant characteristics

| Attribute | YA (n=5) | MAA (n=5) | OA (n=5) |
|--|---------------------------|---------------------------|----------------------------|
| Mean Age & Standard deviation | 21.7±2.45 years | 51.2±7.20 years | 69.1±6.62 years |
| Gender | Male: n=7 Female: n=23 | Male: n=8 Female: n=22 | Male: n=13 Female: n=17 |
| Mean score on Language Experience and Proficiency Questionnaire | 7.83±1.12 | 8.06±0.82 | 7.87±0.591 |
| Mean socioeconomic status score on Modified Kuppuswamy scale | 20.8±4.19 | 21.9±3.71 | 21.6±3.76 |

Caption: YA = Young adults; MAA = Middle-aged adults; OA = Older adults

| Instructions & Prompts | | Technical Aspects | |
|---|--|---|--|
| Participant instruction slide | Response prompt slide | Themes: Home/Family Errands and Work/Professional Errand | |
| You'll see several scenes from everyday life in which a character is given multiple tasks to complete at different timings in a day. You must help the character order the tasks according to the timeline. The number of tasks given to the character increases as the assessment progresses. Carefully attend to each of the tasks. When the screen flashes 'Help me order,' please respond within 60 seconds. Accuracy is more important than speed; we will evaluate your response based on its correctness, not how quickly you complete it. Remember, the maximum response time is 60 seconds. Press the 'Proceed' button after completing your response. | The concluding response prompt slide for each trial instructs participants with "help me order", indicating their goal of correctly sequencing the activities. | Characters: Family member, Working Parent, Homemaker, Policeman, Vegetable vendor, Railway officer, Gardener, Postman, Doctor, Peon, Bank Manager, Tailor, Milkman, Actor, Chef, Hotel Employee, Professor | |
| | | Setting: Living Room, Dining Hall, Balcony, Garden, Doorway, Stable, Kitchen, School, College, Hospital, Garden, Shoot Location, Hotel, Bank, Office, Vegetable stall, Railway station, Police station, Tailoring Shop | |
| | | Stimuli length of each activity instruction: 15-20 words | |
| | | Response acquisition window: 60 seconds | |
| Example of a 6-span trial token | | | |
| It's 8:00 a.m. now and Mrs. Aysha is at home. Her family/friends will approach her each with a task for her to complete. Carefully attend to each of the tasks! You are supposed to help Mrs. Aysha order the tasks in the sequence that they are to be completed based on the timeline. Give your response when the screen flashes "Help me order." | | | |
| <ol style="list-style-type: none"> 1. Aunt, guests will be arriving for the party tonight. Please blow balloons by 8:30 p.m. 2. Aysha, the gas cylinder is empty. Ask your son to book it by 10:00 a.m. 3. Aysha, our dog is feeling quite uneasy. Please take him for a walk before 6:00 p.m. 4. Aunt, I have put my phone for charging. Please switch it off at 3:00 p.m. 5. Mama, I need fresh flowers to make a bouquet. The florist closes by 9:30 a.m. 6. Darling, the school is ready to give our daughter an admission. Please pay the fees by 11:00 a.m. | | | |
| Scoring | | | |
| WM Span | | Metacognitive Facet | |
| Highest successfully completed span before failing both tokens within a span | | Estimation discrepancy score: difference between predicted/postdicted spans and actual performance | |

Figure 1. Components of the AOT-M

Tasks to assess concurrent validity

Sentence Ordering (SO) task

SO task is a span-based measure designed to assess WM⁽⁶⁾. Participants are required to recall sentences and rearrange the words in increasing order of the word length. This task employs both auditory and visual modalities simultaneously, like AOT-M and presents everyday sentences of varying complexity using SuperLab software. Sentences range from three to ten words, with complexity levels equivalent to the participant's WM span. Each sentence contains words of different lengths in terms of phonemes/letters and syllables, ensuring no two words within a sentence have the same length. Every complexity level, from Level 3 (three-word sentences) to Level 10 (ten-word sentences), is represented by two trials, offering participants a second chance if the initial attempt is unsuccessful with an ISI of 1000 milliseconds. The task is terminated when participants incorrectly perform on two consecutive trials. The WM span is recorded as the highest level at which participants can accurately order words in the target sentence stimuli in ascending order of word length. For instance, if a participant successfully orders a 6-word sentence but fails on two trials of a 7-word sentence, their WM span is recorded as 6. The SO task possesses optimum psychometric properties, including moderate test-retest reliability (Intraclass Correlation Coefficient = 0.559) and strong concurrent validity ($r = 0.623, p < 0.001$)⁽⁶⁾.

Digit Letter Ordering (DLO) task

It is an adapted version of the Letter-Number Sequencing task from the Wechsler Adult Intelligence Scale-IV⁽²⁹⁾, designed to measure WM⁽⁶⁾. In this task, participants are presented with a series of letters and digits through both auditory and visual modalities. They are required to first arrange the letters in alphabetical order followed by the digits in ascending numerical order. The task consists of levels ranging from 3 items to 10 items, increasing in complexity. Each stimulus is presented for 2000 ms, with an ISI of 1000 ms. The task is terminated when participants perform incorrectly on two consecutive trials. The WM span is recorded as the highest level up to which the participants can accurately order the items. For instance, if a participant successfully orders a 5-item set of digits and letters but fails on two trials of the 6-item set, their performance span is recorded as 5. The DLO task possesses optimum psychometric properties, including test-retest reliability (Intraclass Correlation Coefficient = 0.619) and strong concurrent validity ($r = 0.634, p < 0.001$)⁽⁶⁾.

Procedure

Initially, information was collected from participants to gather demographic data and ascertain the inclusion criteria. Participants were informed about the study and requested to provide the signed consent form. Participants were then informed that three tasks would be administered: AOT-M, SO and DLO. Task instructions explaining the nature of each task were provided, followed by a

maximum of two practice trials to allow for task familiarization. Following task familiarization using practice trials, participants predicted their performance using a Likert scale (1-5), rating their confidence in completing each span. Spans that received ratings of 4 and 5 were considered prediction spans. For example, a rating of 4 for the sixth span in the SO task indicated a prediction span of 6. After making predictions, participants performed the tasks one after the other in random order. WM performance span was assessed based on the highest span correctly ordered by participants in each task. WM performance span was determined by the number of activities correctly ordered on the AOT-M, the number of words correctly ordered on the SO and the number of digits & letters correctly ordered on the DLO. Postdiction ratings were taken using similar questions as the predictions immediately after performing each task. Prediction and postdiction questions are outlined in Table 2. The metacognitive assessment across all three tasks was determined by calculating the estimation discrepancy, which is the difference between the predicted or postdicted spans and the actual task performance span. AOT-M, including its metacognitive facet, was readministered on 10 participants from each age group after a ten-day interval to evaluate test-retest reliability. The sequence of the entire procedure has been outlined in the Figure 2.

Statistical analysis

Statistical analysis was done using SPSS version 26. Descriptive statistics were utilized to examine the distribution of performance

spans and estimation discrepancies across various age groups. The Independent-Samples Kruskal-Wallis Test was employed to compare performance spans and estimation discrepancies between the groups. The Chi-square test was used to evaluate differences in gender distribution across the three age groups, while the Mann-Whitney U test was employed to explore the effect of gender on performance spans and estimation discrepancies. The Wilcoxon Signed Ranks Test was used to assess differences between prediction and postdiction estimation discrepancy values. The Kolmogorov-Smirnov test was employed to assess the normality of the data.

Reliability

Test-retest reliability was assessed using the Intraclass Correlation Coefficient⁽²³⁾. Bland Altman plots were constructed to assess the systematic error (mean difference) and the 95% limits of agreement between the AOT-M initial and subsequent assessment results.

Validity

Ecological validity

The ecological validity was assessed using the Multidimensional Assessment of Research in Context (MARC) tool⁽³⁰⁾, which evaluates the extent to which psychological and neuroscientific studies capture real-world behavior. MARC tool enables researchers to explicitly report the level of ecological validity by answering seven questions about the study's design, tasks, stimuli, measures,

Table 2. Prediction-postdiction probes

| Prediction question | Task | Postdiction question |
|---|---|--|
| Up to how many activities do you think you can order correctly? | AOT-M | Up to how many activities do you think you have ordered correctly? |
| Up to how many word sentences do you think you can order correctly? | SO (George et al., 2020 ⁽⁶⁾) | Up to how many word sentences do you think you have ordered correctly? |
| Up to how many digits & letters do you think you can order correctly? | DLO (George et al., 2020 ⁽⁶⁾) | Up to how many digits & letters do you think you have ordered correctly? |

Caption: AOT-M = Activity ordering task with metacognitive facet; SO = Sentence ordering task; DLO = Digit Letter Ordering task

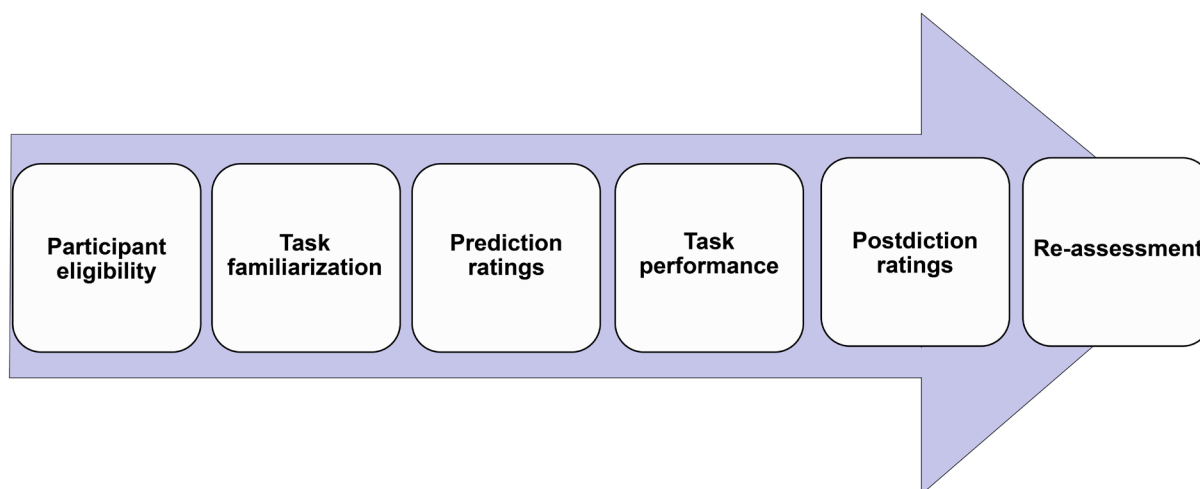


Figure 2. Procedure outline

participant sampling and stakeholder involvement. The tool provides a compass plot that visually represents the balance among controlled, partially naturalistic and naturalistic approaches.

Construct validity

The construct validity of the WM component of AOT-M was assessed through Rasch analysis to determine whether the task measures the intended construct of everyday WM. Additionally, the analysis aimed to ascertain if the difficulty of the AOT-M increases as task levels progress and whether these levels effectively differentiate individuals with varying WM capacities. A Wright map generated through Rasch analysis allows researchers to visually compare the predicted order of item difficulty with the actual order observed in the dataset.

Concurrent validity

The concurrent validity of the AOT-M to measure one's WM span was assessed by the calculation of the Pearson's correlation coefficient between the AOT-M and the SO task as well as the DLO task as reference measures. Concurrent validity of the metacognitive facet of AOT-M was assessed by the Pearson's correlation coefficient between the estimation discrepancies on AOT-M and the SO task as well as the DLO task as reference measures. Concurrent validity was assessed using Pearson's correlation coefficient⁽³¹⁾. Bland Altman plots were constructed to assess the systematic error (mean difference) and the 95% limits of agreement between the performance spans on AOT-M and DLO & SO.

RESULTS

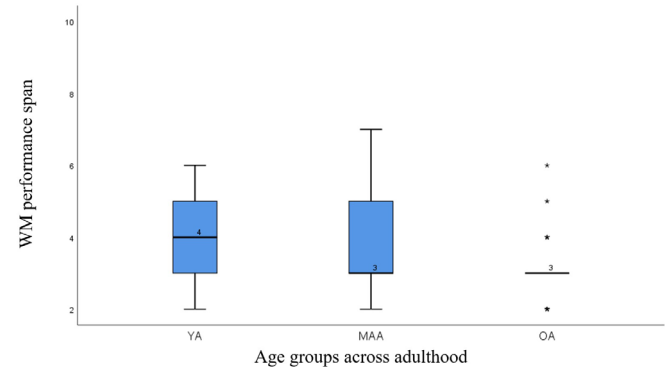
Age related cognitive changes on AOT-M

YA demonstrated the highest performance spans, with a median of 4.00 and interquartile ranges (IQR) spanning from 3.00 (Q1) to 5.00 (Q3) as compared to MAA who showed a slight decline, with a median of 3.00 and IQR values ranging from 3.00 (Q1) to 5.00 (Q3). OA exhibited the lowest performance spans, with both the median and Q3 values at 3.00 and Q1 at 2.75. The Kruskal-

Wallis test revealed significant differences in performance spans across the three age groups ($H = 11.002, p = 0.004$). Subsequent Dunn's pairwise comparisons highlighted significant differences between MAA and OA ($H = 14.217, p = 0.023$) and between YA and OAA ($H = 20.183, p = 0.001$), while the difference between YA and MAA was not significant ($H = 5.967, p = 0.34$). These findings demonstrate that the AOT-M seems to be sensitive to age-related declines in WM, effectively tapping into cognitive changes associated with aging. The median performance span on AOT-M across various age groups are displayed in the Figure 3. Median and IQR of performance spans on AOT-M are given in Table 3. Chi-square analysis revealed no statistically significant difference in gender distribution across the three age groups ($\chi^2 (2, N = 90) = 3.21, p = 0.2$). Additionally, no significant gender effect was found on the performance spans ($U = 699, p = 0.112$). Females exhibited a median performance span of 3.00 (IQR: 3.00 to 5.00), while males reported a median performance span of 3.00 (IQR: 3.00 to 3.75).

Age related metacognitive changes on AOT-M

The estimation discrepancies on the prediction/postdiction spans (prediction/postdiction span-performance span) on AOT-M were compared between groups:



Caption: AOT-M = Activity ordering task with metacognitive facet; YA = Young adults; MAA = Middle-aged adults; OA = Older adults; WM= Working memory; * indicates outliers

Figure 3. Median performance span on AOT-M across age groups

Table 3. Median and interquartile ranges of performance span, prediction and postdiction estimation discrepancy values on AOT-M

| Variable | Age Group | Q1 | Median | Q3 | Kruskal-Wallis H statistic | p-value | Comparison group | Kruskal-Wallis H statistic | p-value | | |
|---|-----------|-------|--------|------|----------------------------|---------|------------------|----------------------------|---------|--------|--------|
| AOT-M performance span | YA | 3 | 4 | 5 | 11.002 | 0.004* | YA vs MAA | 5.967 | 0.34 | | |
| | MAA | 3 | 3 | 5 | | | MAA vs OA | | | 14.217 | 0.023* |
| | OA | 2.75 | 3 | 3 | | | YA vs OA | | | 20.183 | 0.001* |
| AOT-M prediction estimation discrepancy | YA | 0 | 2 | 3 | 7.148 | 0.028* | YA vs MAA | -17.25 | 0.009* | | |
| | MAA | 2 | 3 | 4.25 | | | MAA vs OA | | | 5.1 | 0.442 |
| | OA | 1.75 | 2.5 | 3 | | | YA vs OA | | | -12.15 | 0.067 |
| AOT-M postdiction estimation discrepancy | YA | -1 | 0 | 1 | 0.851 | 0.654 | | | | | |
| | MAA | -0.25 | 0 | 1 | | | | | | | |
| | OA | 0 | 0 | 1 | | | | | | | |

*Indicates statistically significant p-value ($p < 0.05$)

Caption: AOT-M = Activity ordering task with metacognitive facet; YA = Young adults; MAA = Middle-aged adults; OA = Older adults;

Prediction Estimation Discrepancy: Significant differences were found in prediction estimation discrepancies across age groups ($H = 7.148, p = 0.028$). Dunn’s pairwise tests revealed significant differences between YA and MAA ($H = -17.250, p = 0.009$). YA showed a median discrepancy of 2.00 (IQR: 0.00 to 3.00), MAA exhibited a higher median discrepancy of 3.00 (IQR: 2.00 to 4.25) and OA had a median discrepancy of 2.50 (IQR: 1.75 to 3.00). These results indicate that MAA & OA tend to misestimate their performance more than YA, highlighting age-related differences in prediction estimation accuracy on AOT-M. No significant gender effect was found on prediction estimation discrepancies ($U = 850, p = 0.873$). Females exhibited a median prediction estimation discrepancy of 2.00 (IQR: 1.00 to 3.25), whereas males demonstrated a median prediction estimation discrepancy of 2.00 (IQR: 2.00 to 3.00);

Postdiction Estimation Discrepancy: No significant differences were observed in postdiction estimation discrepancies across age groups ($H = 0.851, p = 0.654$). YA had a median discrepancy of 0.00 (IQR: -1.00 to 1.00), MAA had a median discrepancy of 0.00 (IQR: -0.25 to 1.00) and OA showed a median discrepancy of 0.00 (IQR: 0.00 to 1.00). Median and IQR of prediction and postdiction estimation discrepancies on AOT-M are given in Table 3. No significant gender effect was found on postdiction estimation discrepancies ($U = 699, p = 0.112$) ($U = 850, p = 0.873$) ($U = 857.5, p = 0.922$). Both females and males demonstrated a median postdiction estimation discrepancy of 0.00 (IQR: 0.00 to 1.00).

Reliability

The *test-retest reliability* of the AOT-M was evaluated by readministering on 10 participants from each age group after a ten-day interval. The results indicated excellent reliability for AOT-M performance span (Intraclass Correlation Coefficient = 0.966), good reliability for prediction estimation discrepancy (Intraclass Correlation Coefficient = 0.739) and postdiction estimation discrepancy (Intraclass Correlation Coefficient = 0.454). The Bland-Altman plot, illustrating the mean difference (systematic error) and 95% limits of agreement between the initial and subsequent assessment of AOT-M performance spans, is provided as Figure A in the Supplementary Material. Standardized administration and scoring procedures were followed to ensure consistent task execution, including the initiation, termination and scoring of the WM span, as well as the assessment of the metacognitive facet. Therefore, inter-rater reliability was not assessed.

Ecological validity

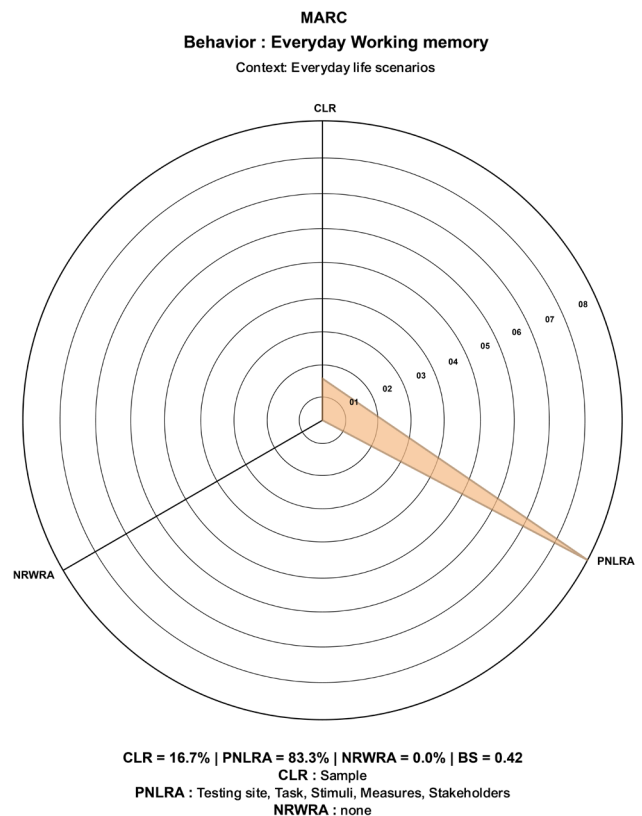
The ecological validity was assessed using the MARC tool which generated a compass plot with the following results: 16.7% Controlled Laboratory Research, 83.3% Partially Naturalistic Laboratory Research Approach, and 0.0% Naturalistic Real-World Research Approach. The balance score was calculated to be 0.42, indicating a predominant focus on the partially naturalistic

laboratory research approach. The generated compass plot is illustrated in Figure 4.

Construct validity

The Rasch analysis was conducted to determine the construct validity of the AOT-M by verifying whether it accurately measures the intended construct of everyday WM and by assessing the increase in difficulty of test items as the levels progress.

The mean of the absolute values of the centered Q3 statistic (MADaQ3) was 0.0712, with a corresponding p-value of 0.355, indicating a good fit to the Rasch model and ascertaining that the AOT-M measures the intended construct. The Q3 correlations, which evaluated the independence of the test items within the AOT-M, revealed low correlations among items, indicating that each test item appears to be distinct. All items had infit values within the acceptable range (0.50 to 1.50)⁽³²⁾, indicating that they fit the model well. While the outfit values for items corresponding to spans 3 and 4 were within the acceptable range, items corresponding to spans greater than 4 had notably low values. These low outfit values for higher span items potentially limit the task’s ability to differentiate between higher levels of WM capacity. The item statistics revealed a progressive range of measures, from -4.095 to 7.39, indicating that the difficulty of test items increased systematically as levels advanced from



Caption: CLR = Controlled laboratory research; PNLRA = Partially naturalistic laboratory research approach; NRWRA = Naturalistic real-world research approach; BS = Balance score

Figure 4. Compass Plot Illustrating Ecological Validity Assessment Using the MARC Tool

Table 4. Item Statistics

| Test item | WM span | Proportion | Measure | S.E. Measure | Infit | Outfit |
|-----------|---------|------------|---------|--------------|-------|--------|
| Item 3 | 3 | 0.8889 | -4.095 | 0.423 | 0.979 | 0.613 |
| Item 4 | 4 | 0.4111 | 0.583 | 0.315 | 0.811 | 0.610 |
| Item 5 | 5 | 0.2667 | 1.996 | 0.348 | 0.693 | 0.373 |
| Item 6 | 6 | 0.1333 | 3.731 | 0.423 | 0.723 | 0.273 |
| Item 7 | 7 | 0.0556 | 5.326 | 0.557 | 0.766 | 0.174 |
| Item 8 | 8 | 0.0111 | 7.39 | 1.057 | 0.947 | 0.104 |

Caption: S.E. = standard error; Infit = Information-weighted mean square statistic; Outfit = Outlier-sensitive means square statistic

span 3 to span 8, as shown in Table 4. This progressive variability in item difficulty can also be appreciated in the Wright map presented in Figure 5. It provides a visual representation of the distribution of respondent latent traits and item difficulties. The left panel of the map shows the distribution of participants' WM performance spans, while the right panel illustrates the item difficulty levels of the various spans. The map indicates that the difficulty level of the spans increases as expected, as fewer participants achieve higher spans. An item reliability of 0.984 was obtained indicating an excellent level of consistency in the item hierarchy however, the person reliability coefficient of 0.606 indicated moderate but inadequate reliability, falling below the accepted threshold of 0.8⁽³²⁾.

Test items 2, 9, and 10 were excluded from the Rasch analysis. Item 2 had uniform responses from all participants, indicating it did not differentiate between different abilities in a dichotomous model. Items 9 and 10 were excluded because no participant could complete them, suggesting they were either too easy or too difficult for the sample, and thus did not contribute meaningful information to the model.

Concurrent validity

Concurrent validity of the performance span and metacognitive facet of the AOT-M was assessed using Pearson's correlation coefficients with the SO and DLO tasks. Moderate-strong significant positive correlations were found between the WM span obtained on the AOT-M and SO tasks (Overall: $r = 0.410$, $p < 0.001$; YA: $r = 0.644$, $p < 0.001$; MAA: $r = 0.407$, $p = 0.019$; OA: $r = 0.545$, $p = 0.001$). Similarly, moderately significant positive correlations were found between performance spans on AOT-M and DLO scores across all age groups (Overall: $r = 0.412$, $p < 0.001$; YA: $r = 0.423$, $p = 0.014$; MAA: $r = 0.502$, $p = 0.003$; OA: $r = 0.695$, $p < 0.001$).

Significant moderate-strong positive correlations were observed between prediction estimation discrepancies on AOT-M and SO across all age groups (Overall: $r = 0.615$, $p < 0.001$; YA: $r = 0.541$, $p = 0.002$; MAA: $r = 0.591$, $p < 0.001$; OA: $r = 0.701$, $p < 0.001$). However, correlations between postdiction estimation discrepancies on AOT-M and SO tasks showed mixed results, with significant correlations found in the YA and MAA groups (YA: $r = 0.280$, $p = 0.008$; MAA: $r = 0.390$, $p = 0.033$), while no significant correlation was observed in the OA group ($r = 0.144$, $p = 0.449$). No significant correlations between prediction/postdiction estimation discrepancies on AOT-M and DLO were obtained. The median and IQR of prediction and

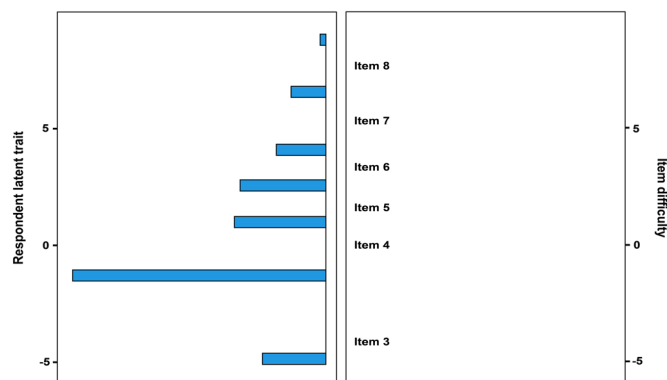


Figure 5. Wright Map illustrating respondent latent trait (WM performance spans) and corresponding item difficulty levels

postdiction estimation discrepancy values on the DLO and SO tasks are provided in Table A in the Supplementary Material. The Bland-Altman plots illustrating the mean difference (systematic error) and 95% limits of agreement between AOT-M and SO (Figure B), as well as AOT-M and DLO (Figure C) are provided in the Supplementary Material.

DISCUSSION

The present study aimed to provide preliminary evidence on the psychometric properties of the AOT-M. Testing a cognitive task in a neurotypical sample is crucial for assessing feasibility and validity thereby, establishing a necessary foundation before its application to clinical populations⁽³³⁾. Therefore, as an initial step, the AOT-M was administered to a non-clinical sample across adulthood.

The age-related trends on the performance spans on AOT-M suggested that the novel task could examine the age-related differences. The decline in performance spans with age aligns with literature indicating age-related declines in WM⁽³⁴⁾. As individuals age, increased cognitive effort may be necessary for task engagement, leading to faster depletion of cognitive resources⁽³⁵⁾. This depletion likely contributes to the decline in performance spans observed on the AOT-M, highlighting its sensitivity to age-related cognitive changes.

The observed pattern of prediction estimation discrepancies across age groups reveals a significant decline in prediction accuracy with age. As individuals transition from young adulthood to middle adulthood, the tendency to misestimate cognitive abilities increases⁽³⁶⁾. The trends from the present study indicate that the

extent of misestimation is higher among MAA and OA compared to YA. These findings align with existing literature that highlights age-related declines in metacognitive accuracy, metacognitive sensitivity and efficiency and increases in metacognitive bias^(14,35). The consistent stability of postdiction estimation discrepancies across adulthood aligns with findings indicating preserved metacognitive abilities in aging on retrospective measures⁽³⁷⁾. Lower estimation discrepancies on postdictions compared to predictions may result from better awareness of performance, enabling more precise calibration. This observation is consistent with literature suggesting that retrospective measures of metacognition tend to be more accurate than prospective measures⁽³⁸⁾. The decline in prediction accuracy and consistent postdiction estimation discrepancies across adulthood observed in AOT-M highlight its potential to reflect age-related changes in metacognitive abilities.

Studies on gender differences in WM present mixed findings, with some reporting significant effects⁽³⁹⁾ and others finding no such differences⁽⁴⁰⁾. Although the present study did not observe gender effect, drawing meaningful conclusions about gender differences may be constrained by the uneven gender representation of participants, as this was not the primary focus of the research.

Understanding the critical need to assess the stability of cognitive assessments designed for repeated use over time⁽²²⁾, the study evaluated the test-retest reliability of the AOT-M. The results indicate good test-retest reliability indicating consistency and stability over repeated administrations for both the WM component and the metacognitive facet. The Bland-Altman plots demonstrated a reasonable alignment between the WM performance spans across repeated administrations, further reinforcing the test-retest reliability of the AOT-M.

The compass plot and balance score from the MARC tool demonstrate that the AOT-M predominantly adopts a partially naturalistic laboratory approach to simulate everyday WM demands. Stimuli comprising of culturally relevant voice-overs in animated videos depicting scenarios like Home/Family Errands and Work/Professional Errands enhance participant engagement and realism. The task's design, which requires participants to prioritize and order activities based on timelines, mirrors the complex WM demands encountered in daily life. Stakeholder involvement across conceptualization, design and implementation phases further validates the task's relevance and applicability to everyday contexts. Taken together, these elements highlight the AOT-M's capacity to effectively replicate everyday scenarios, substantiating its ecological validity.

Rasch analysis was employed to assess the construct validity of the WM component in AOT-M, providing essential indicators to determine whether the task accurately measures the intended construct and aligns with theoretical predictions. The MADaQ3 value was low, with a non-significant p-value, indicating a good fit between the AOT-M data and the Rasch model⁽⁴¹⁾. The Q3 correlations evaluated the independence of items within the AOT-M. Low Q3 correlations among items indicated that each item measures distinct aspects of WM processes independently. The absence of significant misfit and the presence of low item correlations collectively demonstrate that the AOT-M is psychometrically sound in terms of model fit and item independence.

The difficulty of the test items varied in difficulty, ranging from very easy (e.g., Item 3 with a Measure of -4.095) to very difficult (e.g., Item 8 with a Measure of 7.390). Infit values, which assess how well responses to appropriately challenging items align with the Rasch model, generally fall within the optimal range of 0.5 to 1.50⁽³²⁾, indicating a good fit. Conversely, outfit values consistently below the acceptable range across the test items suggest potential overfitting, where items fit the model better than expected. This overfit appears to be linked to an observed performance plateau, where majority of the participants failed to advance beyond initial task levels. Consequently, responses concentrated at lower difficulty levels reduced variability in participant performance and causing the statistical model to closely fit the data at these easier levels.

The Wright map demonstrated a gradual increase in test item difficulty levels, yet it reveals a notable number of respondents achieving spans of 3 and 4. This suggests that while the items are well-ordered in terms of difficulty, there may be a need to refine the task to better differentiate higher levels of WM capacity. The AOT-M task demonstrates a high item reliability value of 0.984, indicating well-defined and stable item difficulty. However, the person reliability coefficient of 0.606 fell below the accepted threshold of 0.8⁽³²⁾, suggesting inadequate consistency in distinguishing individuals based on their WM capacities. A performance plateau was observed at the initial levels of the task, likely attributable to the demanding nature imposed by the brief ISI of 1000 ms and the recall response format. Despite adaptation from established WM tasks^(6,7), this ISI might not have fully met the AOT-M's unique demands, affecting participant performance. Moreover, the recall format of the task could have imposed higher demands than recognition tasks, potentially depleting cognitive resources and impairing WM processing⁽⁴⁾. The reduced variability in WM spans among participants might have led to clustering around similar performance levels, thereby creating the impression of more items than participants and consequently contributing to the observed lower person reliability.

The results of the current study support the concurrent validity of the AOT-M, especially its WM component. Performance spans on the AOT-M showed moderate to strong significant positive correlations with both the DLO and SO tasks across all age groups, confirming that the AOT-M effectively measures WM capacity as intended. The Bland-Altman plots indicated a reasonable alignment between the WM span on AOT-M and both the DLO & SO tasks, further reaffirming the concurrent validity of the AOT-M's WM component. These findings align with established literature indicating that tasks assessing similar constructs typically yield homogeneous results and exhibit strong correlations, thereby demonstrating good concurrent validity⁽⁴²⁾.

Regarding the metacognitive facet of the AOT-M, the prediction estimation discrepancies exhibited moderate to strong correlations with the SO task across all age groups. This indicates that participants' ability to predict their performance on the AOT-M aligned with their predictions on the SO task, reinforcing the concurrent validity of the AOT-M in assessing metacognitive abilities. However, no significant correlations were found between prediction estimation discrepancies on the AOT-M with the DLO task. This lack of correlation could be due to the differences in task nature; while DLO involves remembering and ordering finite

digits and letters, SO and AOT-M are more similar in requiring participants to remember strings of linguistic items and order them based on certain criteria. Specifically, the SO task involves remembering words and ordering them by length, analogous to the AOT-M's requirement of remembering activities and ordering them by timelines. Additionally, no significant correlations were observed between postdiction estimation discrepancies on the AOT-M and either the DLO or SO tasks. This may be attributed to the generally low performance spans on the AOT-M, leading to sharper calibrations in postdiction estimates as participants become more familiar with the task, reducing variability and resulting in weaker correlations.

This study represents an initial exploration of the psychometric properties of the AOT-M, a novel context-based task designed to assess everyday WM with a metacognitive facet. The observed patterns of WM spans and estimation discrepancies across different adult age groups suggest that the task is sensitive to the cognitive and metacognitive changes related to aging. The MARC tool provided evidence for its ecological validity, emphasizing its capacity to replicate everyday WM demands. Rasch analysis supported the construct validity of the AOT-M, confirming its alignment with theoretical expectations and efficacy in assessing WM. However, the moderate person reliability, possibly influenced by the performance plateau attributed to unforeseen cognitive overload from minor factors in the task design, emphasizing the need for future research to optimize these design elements to improve task sensitivity. Concurrent validity was demonstrated through significant correlations with established tasks such as DLO and SO, validating the AOT-M's ability to assess both WM and metacognitive abilities through performance estimation discrepancies. Moreover, the test-retest reliability demonstrated consistent performance across repeated administrations, ensuring its stability for longitudinal use.

Limitations and future directions

The present study was conducted on a small sample of neurotypical adults across adulthood, limiting the generalizability of findings. Future research should prioritize expanding the sample size to encompass broader age ranges and diverse demographic profiles, ensuring a more representative participant pool that includes balanced representation from various professional and academic backgrounds. Additionally, future studies could aim for adequate gender representation to more effectively identify any potential gender effects on various measures of AOT-M. Employing stratified sampling techniques in future studies could mitigate potential biases introduced by convenience sampling, thereby enhancing the study's external validity. Minor task design factors, such as the brief ISI and response format, were identified as potential contributors to increased cognitive demands, resulting in a performance plateau at initial task levels. Future research should focus on refining these task parameters to optimize cognitive load management and improve task sensitivity. Additionally, future studies could extend the psychometric evaluation of the refined task to establish age-specific normative data for WM spans and associated metacognitive facets across different demographics and clinical populations.

CONCLUSION

The present study marks the first step in the comprehensive series of investigations aimed at establishing the psychometric properties of the AOT-M, a novel task designed to assess everyday WM with a metacognitive facet. The study's findings reveal discernible patterns in WM spans and estimation discrepancies across various adult age groups, indicating the task's sensitivity to cognitive and metacognitive changes associated with aging. Preliminary evidence from this study supports the task's ecological and concurrent validity, as well as its test-retest reliability. While Rasch analysis supports its construct validity in measuring WM, the observed moderate person reliability value indicates minor limitations in the task sensitivity. Future research would focus on further refining the AOT-M and establishing its psychometric properties across diverse neurotypical and clinical populations, ensuring a comprehensive and representative assessment of its utility.

ACKNOWLEDGEMENTS

The study is supported by funding received from the Department of Science and Technology (DST) under the Cognitive Science Research Initiative (CSRI), Government of India (DST/CSRI/2018/29).

REFERENCES

1. Muñoz-Pradas R, Díaz-Palacios M, Rodríguez-Martínez EI, Gómez CM. Order of maturation of the components of the working memory from childhood to emerging adulthood. *Curr Res Behav Sci.* 2021;2:100062. <http://doi.org/10.1016/j.crbeha.2021.100062>.
2. Brehmer Y, Westerberg H, Bäckman L. Working-memory training in younger and older adults: training gains, transfer, and maintenance. *Front Hum Neurosci.* 2012;6:63. <http://doi.org/10.3389/fnhum.2012.00063>. PMID:22470330.
3. Pliatsikas C, Verissimo J, Babcock L, Pullman MY, Gleib DA, Weinstein M, et al. Working memory in older adults declines with age, but is modulated by sex and education. *Q J Exp Psychol.* 2019;72(6):1308-27. <http://doi.org/10.1177/1747021818791994>. PMID:30012055.
4. Mohapatra B, Laures-Gore J. Moving toward accurate assessment of working memory in adults with neurogenically based communication disorders. *Am J Speech Lang Pathol.* 2021;30(3):1292-300. http://doi.org/10.1044/2021_AJSLP-20-00305. PMID:33970679.
5. Nigam R, Kar BR. Cognitive ageing in developing societies: an overview and a cross-sectional study on young, middle-aged and older adults in the Indian context. *Psychol Dev Soc J.* 2020;32(2):278-307. <http://doi.org/10.1177/0971333620937511>.
6. George VM, Bajaj G, Bhat JS. Efficacy of working memory training in middle-aged adults. *Commun Sci Disord.* 2020;25(4):830-56. <http://doi.org/10.12963/csd.20768>.
7. Redick TS, Broadway JM, Meier ME, Kuriakose PS, Unsworth N, Kane MJ, et al. Measuring working memory capacity with automated complex span tasks. *Eur J Psychol Assess.* 2012;28(3):164-71. <http://doi.org/10.1027/1015-5759/a000123>.
8. Carrigan N, Barkus E. A systematic review of cognitive failures in daily life: healthy populations. *Neurosci Biobehav Rev.* 2016;63:29-42. <http://doi.org/10.1016/j.neubiorev.2016.01.010>. PMID:26835660.
9. Brown KD, Schmitter-Edgecombe M. A clinic-based measure of everyday planning ability: the overnight trip task. *Arch Clin Neuropsychol.* 2024;39(1):51-64. <http://doi.org/10.1093/arclin/acad052>. PMID:37489707.

10. Okahashi S, Seki K, Nagano A, Luo Z, Kojima M, Futaki T. A virtual shopping test for realistic assessment of cognitive function. *J Neuroeng Rehabil*. 2013;10(1):59. <http://doi.org/10.1186/1743-0003-10-59>. PMID:23777412.
11. Sanz Simon S, Ben-Eliezer D, Pondikos M, Stern Y, Gopher D. Feasibility and acceptability of a new web-based cognitive training platform for cognitively healthy older adults: the breakfast task. *Pilot Feasibility Stud*. 2023;9(1):136. <http://doi.org/10.1186/s40814-023-01359-2>. PMID:37542331.
12. Quiles C, Prouteau A, Verdoux H. Assessing metacognition during or after basic-level and high-level cognitive tasks? A comparative study in a non-clinical sample. *Encephale*. 2020;46(1):3-6. <http://doi.org/10.1016/j.encep.2019.05.007>. PMID:31227210.
13. Woelfer SW, Tomitch LM, Procailo L. Working memory, metacognition and foreign language reading comprehension: a bibliographical review. *Let Hoje*. 2019;54(2):263-80. <http://doi.org/10.15448/1984-7726.2019.2.32314>.
14. Filippi R, Ceccolini A, Periche-Tomas E, Bright P. Developmental trajectories of metacognitive processing and executive function from childhood to older age. *Q J Exp Psychol*. 2020;73(11):1757-73. <http://doi.org/10.1177/1747021820931096>. PMID:32419614.
15. Gilbert SJ, Bird A, Carpenter JM, Fleming SM, Sachdeva C, Tsai PC. Optimal use of reminders: Metacognition, effort, and cognitive offloading. *J Exp Psychol Gen*. 2020;149(3):501-17. <http://doi.org/10.1037/xge0000652>. PMID:31448938.
16. Conte N, Fairfield B, Padulo C, Pelegrina S. Metacognition in working memory: confidence judgments during an N-back task. *Conscious Cogn*. 2023;111:103522. <http://doi.org/10.1016/j.concog.2023.103522>. PMID:37087901.
17. Fulton EK. How well do you think you summarize? Metacomprehension Accuracy in younger and older adults. *J Gerontol B Psychol Sci Soc Sci*. 2021;76(4):732-40. <http://doi.org/10.1093/geronb/gbz142>. PMID:31677351.
18. Arora C, Frantz C, Toglia J. Awareness of performance on a functional cognitive performance-based assessment across the adult lifespan. *Front Psychol*. 2021;12:753016. <http://doi.org/10.3389/fpsyg.2021.753016>. PMID:34803834.
19. Jacob NL, Bajaj G, Rooha A, George VM, Bhat JS. Activity ordering task: conceptualization and development of a novel context-based working memory task with a metacognitive facet. *CoDAS*. 2024;36(6):e20240041. <http://doi.org/10.1590/2317-1782/20242024041en>.
20. Sopian S, Inderawati R, Petrus I. Developing e-learning based local-folklores for eighth graders. *English Rev*. 2019;8(1):101-10. <http://doi.org/10.25134/erjee.v8i1.1813>.
21. Miyake A, Friedman NP. The nature and organization of individual differences in executive functions: four general conclusions. *Curr Dir Psychol Sci*. 2012;21(1):8-14. <http://doi.org/10.1177/0963721411429458>. PMID:22773897.
22. Horan B, Heckenberg R, Maruff P, Wright B. Development of a new virtual reality test of cognition: assessing the test-retest reliability, convergent and ecological validity of CONVIRT. *BMC Psychol*. 2020;8(1):61. <http://doi.org/10.1186/s40359-020-00429-x>. PMID:32532362.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-63. <http://doi.org/10.1016/j.jcm.2016.02.012>. PMID:27330520.
24. Steinborn MB, Langner R, Flehmig HC, Huestegge L. Methodology of performance scoring in the d2 sustained-attention test: cumulative-reliability functions and practical guidelines. *Psychol Assess*. 2018;30(3):339-57. <http://doi.org/10.1037/pas0000482>. PMID:28406669.
25. Oliveira IR, Seixas C, Osório FL, Crippa JAS, Abreu JN, Menezes IG, et al. Evaluation of the psychometric properties of the cognitive distortions questionnaire (CD-Quest) in a sample of undergraduate students. *Innov Clin Neurosci*. 2015;12(7-8):20-7. PMID:26351620.
26. 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(3):189-98. [http://doi.org/10.1016/0022-3956\(75\)90026-6](http://doi.org/10.1016/0022-3956(75)90026-6). PMID:1202204.
27. Radhakrishnan M, Nagaraja SB. Modified Kuppaswamy socioeconomic scale 2023: stratification and updates. *Int J Community Med Public Health*. 2023;10(11):4415-8. <http://doi.org/10.18203/2394-6040.ijcmph20233487>.
28. Marian V, Blumenfeld HK, Kaushanskaya M. The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J Speech Lang Hear Res*. 2007;50(4):940-67. [http://doi.org/10.1044/1092-4388\(2007\)067](http://doi.org/10.1044/1092-4388(2007)067). PMID:17675598.
29. Mielicki MK, Koppel RH, Valencia G, Wiley J. Measuring working memory capacity with the letter-number sequencing task: advantages of visual administration. *Appl Cogn Psychol*. 2018;32(6):745-53. <http://doi.org/10.1002/acp.3468>.
30. Naumann S, Byrne ML, de la Fuente A, Harrewijn A, Nugiel T, Rosen M, et al. Assessing the degree of ecological validity of your study: Introducing the Multidimensional Assessment of Research in Context (MARC). *Mind Brain Educ*. 2022;16(3):188-98. <http://doi.org/10.1111/mbe.12318>.
31. Evans JD. *Straightforward statistics for the behavioral sciences*. Pacific Grove: Thomson Brooks/Cole Publishing Company; 1996. Linear correlation; p. 127-58.
32. Linacre JM. *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago, IL: MESA; 2007.
33. Schuster RM, Mermelstein RJ, Hedeker D. Acceptability and feasibility of a visual working memory task in an ecological momentary assessment paradigm. *Psychol Assess*. 2015;27(4):1463-70. <http://doi.org/10.1037/pas0000138>. PMID:25894710.
34. Devos H, Gustafson K, Ahmadnezhad P, Liao K, Mahnken JD, Brooks WM, et al. Psychometric properties of NASA-TLX and index of cognitive activity as measures of cognitive workload in older adults. *Brain Sci*. 2020;10(12):994. <http://doi.org/10.3390/brainsci10120994>. PMID:33339224.
35. Siegel ALM, Castel AD. Age-related differences in metacognition for memory capacity and selectivity. *Memory*. 2019;27(9):1236-49. <http://doi.org/10.1080/09658211.2019.1645859>. PMID:31339451.
36. D'Souza DF, Bajaj G, George VM, Karuppali S, Bhat JS. "I think I can remember" age-related changes in self-efficacy for short-term memory. *J Nat Sci Biol Med*. 2021;12(1):97. http://doi.org/10.4103/jnsbm.JNSBM_32_20.
37. Mitchell DJ, Cusack R. Visual short-term memory through the lifespan: preserved benefits of context and metacognition. *Psychol Aging*. 2018;33(5):841-54. <http://doi.org/10.1037/pag0000265>. PMID:30091631.
38. Fleming SM, Massoni S, Gajdos T, Vergnaud JC. Metacognition about the past and future: quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neurosci Conscious*. 2016;2016(1):niw018. <http://doi.org/10.1093/nc/niw018>. PMID:30356936.
39. Saylik R, Raman E, Szameitat AJ. Sex differences in emotion recognition and working memory tasks. *Front Psychol*. 2018;9:1072. <http://doi.org/10.3389/fpsyg.2018.01072>. PMID:30008688.
40. Rogala J, Dreszer J, Sińczuk M, Miciuk Ł, Piątkowska-Janko E, Bogorodzki P, et al. Local variation in brain temperature explains gender-specificity of working memory performance. *Front Hum Neurosci*. 2024;18:1398034. <http://doi.org/10.3389/fnhum.2024.1398034>. PMID:39132677.
41. Bazan B. The construction and validation of a new listening span task. *CEFR J Res Pract*. 2021;25(1):39-56. <http://doi.org/10.37546/JALTSIG.TEVAL25.1-4>.
42. Krabbe PFM. *The measurement of health and health status*. San Diego: Academic Press; 2017. Chapter 7, Validity; p. 113-34. <http://doi.org/10.1016/B978-0-12-801504-9.00007-6>.

Author contributions

Nidhi Lal Jacob was responsible for conceptualization, methodology, software, validation, investigation, data curation, formal analysis, writing - original draft, review and editing, resources, visualization, supervision, project administration. *Aysha Rooha* was responsible for conceptualization, methodology, writing - review and editing. *Anjaly S. Nair* was responsible for formal analysis, writing - review and editing. *Gagan Bajaj* was responsible for conceptualization, methodology, software, validation, investigation, formal analysis, writing - review and editing, resources, visualization, supervision, project administration. *Vinitha Mary George* was responsible for project administration, supervision, resources, writing - review and editing. *Jayashree S. Bhat* was responsible for project administration, supervision, writing - review and editing.

SUPPLEMENTARY MATERIAL

Supplementary material accompanies this paper.

Table A. Median and interquartile ranges of prediction and postdiction estimation discrepancy values on Digit Letter Ordering and Sentence Ordering Tasks

Figure A. The Bland-Altman plot illustrating the agreement between the initial and subsequent assessment of AOT-M performance spans

Figure B. The Bland-Altman plot illustrating the agreement between AOT-M performance span and SO performance span

Figure C. The Bland-Altman plots illustrating the agreement between AOT-M performance span and DLO performance span

This material is available as part of the online article from <https://www.scielo.br/j/codas>