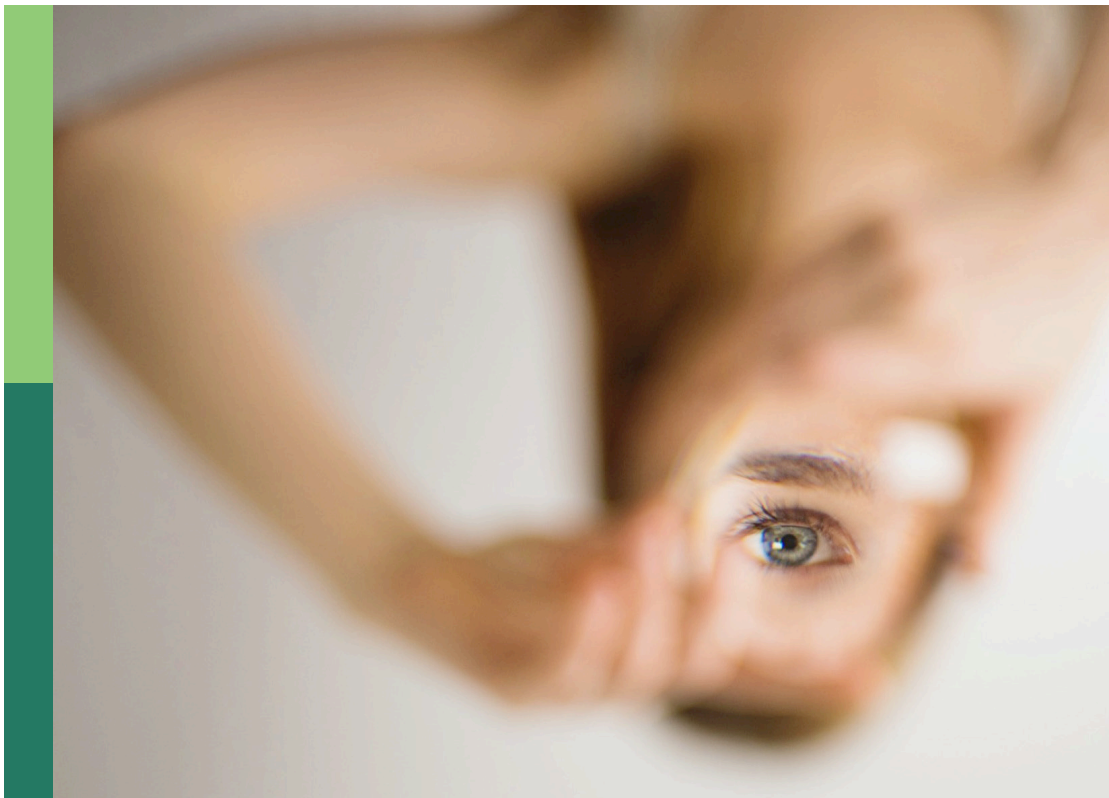
frontiers | Research Topics

Neuropsychological testing: from psychometrics to clinical neuropsychology

Edited by
Alessio Facchin, Elisa Cavicchiolo and Edgar Chan

Published in
Frontiers in Psychology



ISSN 1664-8714
ISBN 978-2-8325-5985-7
DOI 10.3389/978-2-8325-5985-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Neuropsychological testing: from psychometrics to clinical neuropsychology

Topic editors

Alessio Facchin — Magna Graecia University, Italy

Elisa Cavicchiolo — University of Rome Tor Vergata, Italy

Edgar Chan — National Hospital for Neurology and Neurosurgery (NHNN), United Kingdom

Citation

Facchin, A., Cavicchiolo, E., Chan, E., eds. (2025). *Neuropsychological testing: from psychometrics to clinical neuropsychology*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5985-7

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 28 January 2025
DOI 10.3389/fpsyg.2025.1549236

OPEN ACCESS

EDITED AND REVIEWED BY
Pietro Cipresso,
University of Turin, Italy

*CORRESPONDENCE
Alessio Facchin
alessio.facchin@unicz.it

RECEIVED 20 December 2024
ACCEPTED 15 January 2025
PUBLISHED 28 January 2025

CITATION
Facchin A, Cavicchiolo E and Chan E (2025) Editorial: Neuropsychological testing: from psychometrics to clinical neuropsychology. *Front. Psychol.* 16:1549236. doi: 10.3389/fpsyg.2025.1549236

Editorial: Neuropsychological testing: from psychometrics to clinical neuropsychology

Alessio Facchin[1]*, Elisa Cavicchiolo[2] and Edgar Chan[3]

[1]Neuroscience Research Center, Department of Medical and Surgical Sciences, Magna Graecia University, Catanzaro, Italy, [2]Department of Systems Medicine, University of Rome Tor Vergata, Rome, Italy, [3]Department of Neuropsychology, The National Hospital for Neurology and Neurosurgery, London, United Kingdom

KEYWORDS

psychometrics, testing, test development research, reliability, validity

Editorial on the Research Topic
Neuropsychological testing: from psychometrics to clinical neuropsychology

Neuropsychological testing represents an essential part of the clinical examination of neurological patients, and these measures remain the primary instrument for clinical research in neuropsychology (Bauer et al., 2012; Bondi and Smith, 2014; Howieson, 2019). It is crucial that neuropsychological tests are regularly reviewed and updated in order to remain relevant and useful. New research is needed to improve neuropsychological testing as well as help understand the psychometric characteristics and theories behind the tests we use (Bilder and Reise, 2019; Casaletto and Heaton, 2017; Randolph, 2002). This Research Topic on "*Neuropsychological Testing: From Psychometrics to Clinical Neuropsychology*" brings together a collection of articles that examine recent developments in test development and validation across a range of cognitive domains and clinical settings. The emerging picture underlines the complexity of bridging clinical needs with basic psychometric research.

New test development in emerging areas

The development of novel neuropsychological tests is crucial to advance our understanding of brain-behavior relationships in the ever changing social context. Innovative testing methods which incorporate new technology or advances in cognitive neuroscience allow us to better capture cognitive changes and provide more personalized treatment plans (Parsons and Duffield, 2020). As the field of neuropsychology and neurorehabilitation moves toward a greater dependence on computerized or digitalized tools, it is important to consider the suitability of these tools for the individual.

The article (Stoll et al.) explores this concept using the "Digital Tools Test" (DIGI), a standardized instrument designed to evaluate digital tool competencies in a sample of young people and older adults. Preliminary results highlight performance differences between age groups, with older adults showing lower proficiency in navigating digital tools. In the future, digital tool competency assessments like the DIGI may be used in standard neuropsychological assessments. As technological advances allow for biometric measurements to be more accessible, the study (Gomes et al.) explores the use of both response type/time and eye-fixation measures to detect feigned memory

impairment through a computerized version of the well-established Test of Memory Malingering (TOMM). Results found distinct behavioral patterns for genuine and feigned memory impairment. The findings highlight the potential of how eye-tracking metrics may enhance standard paper-and-pencil neuropsychological tools. Finally, the opinion piece (Finley) discusses the use of digital technologies to enhance Performance Validity Assessment (PVA).

Taking an alternative approach, the article (Elkana) explores the "frontal lobe paradox" by discussing the importance of using Real-Life Tasks (RLTs) to enhance standard paper-and-pencil tasks. The "frontal lobe paradox" is a well-described phenomena in neuropsychology whereby some patients with frontal lobe compromise report a host of executive difficulties in daily activities but perform reasonably well in standardized neuropsychological tests. A framework for assessing frontal dysfunction using a variety of RLTs is presented.

Psychometric evaluation or validation

The evaluation of psychometric properties is essential for selecting reliable and valid instruments, making it a fundamental aspect of clinical practice and research in many areas (Souza et al., 2017). Unfortunately, many instruments still lack thorough or complete validation, which hinders their practical application (Monticone et al., 2021). In this Research Topic, particular emphasis has been placed on the psychometric properties of various existing neuropsychological instruments, and notable advancements have also been reported.

The study (de Oliveira et al.) presents the development and initial validation of a new tool for the Assessment of Reading and Executive Functions (AREF) in children. The findings highlight the interdependence of executive functions, such as inhibitory control, cognitive flexibility and working memory, with reading skills. Once new tests such as the AREF are validated and in use, further validation studies and developments can improve its clinical utility. Country-specific validation of tests is useful to overcome inherent cultural, language and educational differences. The study (Taroza et al.) investigated the psychometric properties of the EQ-5D-5L instrument for assessing health-related quality of life (HRQoL) in Lithuanian individuals who have experienced stroke, while the study (Shi and Zhang) investigated the reliability and validity of the Broken ring enVision search (BReViS) test for assessing attention in the Chinese population.

It is also important to understand the test-retest reliability of our tools for monitoring change over time. The study (Isernia et al.) investigates the test-retest reliability of the Yoni-48 task, a tool for assessing Theory of Mind (ToM) in social cognition, and to establish the minimal detectable change for determining clinical significance. Lastly, shortening established tests can often improve clinical utility but it is important that the same validation rigor is applied before use. The study (De Luca et al.) focuses on the development of the Short Italian Wilkins Rate of Reading Test to enhance the test's applicability to elderly and neuropsychological patients by reducing reading time compared to the original standard form.

Reviews

Meta-analysis and systematic reviews provide a comprehensive understanding of test properties by synthesizing vast amounts of research on a given topic. These studies help ascertain clinical utility with greater power and guide future research. The article (Malek-Ahmadi and Nikkhahmanesh) presents a systematic review assessing the diagnostic accuracy of the Montreal Cognitive Assessment (MoCA) for detecting amnestic mild cognitive impairment. The findings support the MoCA's utility as a screening tool in clinical settings but emphasizes the need for context-specific cutoff adjustments. The article (Maiuolo et al.) provides a critical evaluation of the scale used to assess wellbeing in people with Parkinsonism. Although eight HRQoL tools were identified, questions were raised about the psychometric properties of the measures which may mar their utility.

Summary

Articles in the Topic highlight the interplay between psychometrics and Clinical Neuropsychology. Continued research into novel measures, applications, comparisons and updates is crucial for maintaining and improving the clinical practice of neuropsychological testing.

Author contributions

AF: Conceptualization, Writing – original draft, Writing – review & editing. ECa: Writing – original draft, Writing – review & editing. ECh: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The 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.

Publisher's note

References

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., and Naugle, R. I. (2012). Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch. Clin. Neuropsychol.* 27, 362–373. doi: 10.1093/arclin/acs027

Bilder, R. M., and Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here? *Clin. Neuropsychol.* 33, 220–245. doi: 10.1080/13854046.2018.1521993

Bondi, M. W., and Smith, G. E. (2014). Mild cognitive impairment: a concept and diagnostic entity in need of input from neuropsychology. *J. Int. Neuropsychol. Soc.* 20, 129–134. doi: 10.1017/S1355617714000010

Casaletto, K. B., and Heaton, R. K. (2017). Neuropsychological assessment: past and future. *J. Int. Neuropsychol. Soc.* 23, 778–790. doi: 10.1017/S1355617717001060

Howieson, D. (2019). Current limitations of neuropsychological tests and assessment procedures. *Clin. Neuropsychol.* 33, 200–208. doi: 10.1080/13854046.2018.1552762

Monticone, M., Galeoto, G., Berardi, A., and Tofani, M. (2021). "Psychometric properties of assessment tools," in *Measuring Spinal Cord Injury: A Practical Guide of Outcome Measures*, 7–15. doi: 10.1007/978-3-030-68382-5_2

Parsons, T., and Duffield, T. (2020). Paradigm shift toward digital neuropsychology and high-dimensional neuropsychological assessments. *J. Med. Internet Res.* 22:e23777. doi: 10.2196/23777

Randolph, C. (2002). Neuropsychological testing: evolution and emerging trends. *CNS Spectr.* 7, 307–312. doi: 10.1017/S1092852900017727

Souza, A. C. D., Alexandre, N. M. C., and Guirardello, E. D. B. (2017). Psychometric properties in instruments evaluation of reliability and validity. *Epidemiol. Serv. Saude* 26, 649–659. doi: 10.5123/S1679-49742017000300022

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 06 December 2023
DOI 10.3389/fpsyg.2023.1284859

OPEN ACCESS

EDITED BY
Alessio Facchin,
University of Milano-Bicocca, Italy

REVIEWED BY
Isa Zappullo,
University of Campania Luigi Vanvitelli, Italy
Neringa Grigutyte,
Vilnius University, Lithuania

*CORRESPONDENCE
Saulius Taroza
✉ saulius.taroza@lsmuni.lt

RECEIVED 29 August 2023
ACCEPTED 09 November 2023
PUBLISHED 06 December 2023

CITATION
Taroza S, Burkauskas J, Mickuviene N, Kazukauskiene N and Podlipskyte A (2023) Psychometrics and validation of the EQ-5D-5L instrument in individuals with ischemic stroke in Lithuania. *Front. Psychol.* 14:1284859. doi: 10.3389/fpsyg.2023.1284859

Psychometrics and validation of the EQ-5D-5L instrument in individuals with ischemic stroke in Lithuania

Saulius Taroza*, Julius Burkauskas, Narseta Mickuviene, Nijole Kazukauskiene and Aurelija Podlipskyte

Laboratory of Behavioral Medicine, Neuroscience Institute, Lithuanian University of Health Sciences, Palanga, Lithuania

Background: Experiencing stroke is associated with deterioration in health-related quality of life (HRQL). One of the generic tools used for HRQL assessment is the EuroQol instrument of five dimensions and five levels (EQ-5D-5L), which has not yet been validated in Lithuania. This study aimed to evaluate validity, reliability, and factor structure of the EQ-5D-5L instrument in a sample of Lithuanian individuals at the end of the first week after experiencing ischemic stroke (IS).

Methods: The study had a cross-sectional design, including 134 individuals [61.9% men and 38.1% women; median (IQR) age was 66 years (59–73) years, in the final analysis]. Alongside the EQ-5D-5L, psychological distress was evaluated using the Hospital Anxiety and Depression Scale (HADS), Patient Health Questionnaire-9 (PHQ-9), and Generalized Anxiety Disorder Assessment-7 (GAD-7); neurological impairment with the National Institutes of Health Stroke Scale (NIHSS); and functional independence with the Barthel index (BI). Confirmatory factor analysis (CFA) was performed for validation of the factor structure.

Results: The internal consistency of the EQ-5D-5L instrument was 0.81. A significant ceiling effect (17.2%) of the descriptive part of the EQ-5D-5L was detected. The convergent validity of the EQ-5D-5L descriptive system was confirmed, with significant correlations with the other scales used, except for the visual analog scale. The two-factor ("physical" and "emotional") model was confirmed by CFA, with acceptable fit [root mean square error of approximation (RMSEA) = 0.045, RMSEA 90% CI = 0.000–0.145; comparative fit indices (CFI) = 0.996; non-normal fit index (NFI) = 0.983; Tucker–Lewis Index (TLI) = 0.936; $\chi^2/df = 1.27$)].

Conclusion: This study provides information on the psychometric properties of the EQ-5D-5L instrument in Lithuanian individuals, showing that the EQ-5D-5L descriptive system is a reliable and valid tool for HRQL assessment. The Lithuanian version of the descriptive part of the EQ-5D-5L instrument is best expressed as a two-factor model, estimating the physical and emotional dimensions of HRQL in individuals who have experienced IS.

KEYWORDS

psychometrics, quality of life, ischemic stroke, depression, cross-sectional studies, Lithuanian people, anxiety

1 Introduction

HRQL is recognized as of paramount importance in health outcomes (Kaplan, 1990; Bunevicius et al., 2022). In one study, a single question about self-rated health has been shown to be strongly associated with mortality at follow-up (DeSalvo et al., 2006). Although there are some problems arising with the universal definition of HRQL (Karimi and Brazier, 2016), it is usually described as the daily level of functioning and perceived health-associated wellbeing on a personal level (Stenman et al., 2010). Therefore, the multifaceted construct of HRQL is characterized subjectively by an individual as the impact of illness and its treatment on physical, mental, and social domains of functioning (Revicki et al., 2014). It is assumed that the evaluation of HRQL enables better patient-directed healthcare than the traditional biomedical model, which is focused primarily on diagnosis and treatment (Kaplan, 2003).

Although there are many HRQL instruments implemented in practice, according to previous research, there is no "best" or "worst" instrument (Coons et al., 2000); the choice should depend on the purpose of the measurement. The attractiveness of each instrument depends on the ease of use, its psychometric properties, free availability, and usefulness in the economic assessment of public health interventions. One such instrument belongs to one of the most widely used generic methods of HRQL assessment—the EQ-5D set of instruments (Pequeno et al., 2020). The latest version of the EQ-5D for adults is the EQ-5D-5L, which has better psychometric properties (increased reliability and sensitivity) than its precedent, the EQ-5D-3L (Feng et al., 2021). Regarding the psychometric properties of the EQ-5D-5L, this instrument is valid and reliable for health status assessment across a broad spectrum of populations, with acceptable responsiveness. However, it has some limitations, including the tendency for a ceiling effect and the lack of positive health aspects (Feng et al., 2021).

According to the EQ-5D-5L factor structure, at least one study has suggested that it consists of two latent factors encompassing physical and psychological functioning (Gao et al., 2019), but other studies suggested one-factor structure (Bilbao et al., 2022). However, some concern has recently been raised over the scale's lack of social dimension (Chen and Olsen, 2020). Despite the aforementioned limitations, this scale is used widely due to its simplicity, free-of-charge use for non-commercial reasons, availability in many languages, and applicability for various conditions (Lau et al., 2022).

Stroke, as one of the most frequent worldwide causes of disability (Campbell and Khatri, 2020), is associated with reduced post-stroke HRQL (Cadilhac et al., 2010; Gall et al., 2010; Mar et al., 2015; Chen et al., 2019). In the United States, it has been shown that the consequences of stroke significantly impair the HRQL of respondents who are not committed to an institution compared with those without stroke (Xie et al., 2006). Another study, based on a population in northern Manhattan study, showed a significant worsening of HRQL independent of various risk factors, including functional independence, during the 5-year follow-up (Dhamoon et al., 2010). On the contrary, a study conducted in Lithuania with stroke survivors after 3 and 12 months using the 12-item Short Form Survey of Health showed that the survivors had poorer HRQL than the controls but showed remarkable improvement over time (Kranciukaite-Butylkiniene, 2014). Furthermore, hyperacute recanalization therapy in acute ischemic stroke (IS) is not clearly related to better long-term HRQL, despite better functional outcomes (Kainz et al., 2021). Based on the results of the mentioned studies, it is important to continue research on impaired post-stroke HRQL in order to better understand this phenomenon and thus make suggestions for HRQL improvement-directed interventions.

In terms of HRQL for stroke patients, the validity of the EQ-5D-5L instrument was recently demonstrated for individuals from Poland after stroke (Golicki et al., 2015). Another study performed in Taiwan proved the validity of this instrument for HRQL assessment in patients after stroke undergoing rehabilitation (Chen et al., 2016). Furthermore, a systematic review of the instruments for assessing self-reported HRQL after stroke showed that the EQ-5D instrument was the best choice (Cameron and Wales, 2022).

Given that the EQ-5D-5L has not been validated in Lithuania, this study focused on the psychometric properties, including applicability, internal consistency, validity, and factor structure of this instrument in Lithuanian residents who had experienced IS.

2 Materials and methods

2.1 Study procedure

This study was a part of a research described previously (Burkauskas et al., 2014). Individuals who had experienced acute IS and were admitted to the three different Lithuanian health institutions (Klaipeda University Hospital, Hospital of the Lithuanian University of Health Sciences Kauno Klinikos, and Klaipeda Seamen's Hospital) were invited by a neurologist in the emergency room on duty to participate in the study during two 1-year periods, starting in 2013 and 2016, respectively. In total, 612 consecutive individuals were asked to participate in this study.

The inclusion criteria were: (1) ages 18–80 years; (2) current diagnosis of acute IS as described by the World Health Organization criteria (Hatano, 1976), affirmed by neurovisual imaging with brain-computer or magnetic resonance tomography; and (3) capable of communication and cognition, according to a Mini-Mental State Exam (MMSE) score of more than 19, assessed at the end of the first week. The exclusion criteria were: (1) co-diagnosis of severe pathology (infection, liver and/or renal insufficiency, and malignancy); (2) noted thyroidopathy and/or intake of thyroid-affecting substances; and (3) arrival 2 days later after the onset of IS.

The following characteristics of the individuals were assessed in the emergency department: (1) age; (2) sex; (3) body mass index; (4) presence of premorbid disability, defined as dependency in daily activity according to a Modified Rankin Scale (mRS) score of $\geq$3; (5) use of antithrombotic drugs; (6) chemical thrombolysis; and (7) stroke risk factors, including arterial hypertension, atrial fibrillation, smoking, diabetes mellitus, previous cerebral ischemic event, and experienced myocardial infarction. In addition, the

individuals' neurological impairment was assessed using the National Institutes of Health Stroke Scale (NIHSS) (Spilker et al., 1997).

At the end of their hospital stay, all study individuals were asked to fill out questionnaires in paper form: (1) EQ-5D-5L (Herdman et al., 2011) for HRQL and (2) Hospital Anxiety and Depression Scale (HADS) (Zigmond and Snaith, 1983), Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001), and Generalized Anxiety Disorder assessment-7 (GAD-7) (Spitzer et al., 2006) for psychological distress assessment. At this time point after IS, individuals were checked for functional independence according to the Barthel index (BI) (Mahoney and Barthel, 1965). At the end of the first year of the study, participants were asked to fill in the EQ-5D-5L once more.

For a sufficient sample, power analysis was based on the suggested rule—at least 10 respondents to 1 scale item (Boateng et al., 2018), and it was more than 50 individuals in our case. Figure 1 shows the selection of individuals for the study. In total, 134 individuals were included in the final analysis.

The study was conducted in accordance with the Declaration of Helsinki and met the requirements of the Regional Biomedical Research Ethics Committee, with the assigned licenses P1-BE-2-11/2013 and P2-BE-2-11/2013. Individuals were included only after giving written consent for participation in this study.

2.2 Measurements and applied questionnaires

2.2.1 Modified rankin scale

This global disability-assessing instrument is used to evaluate dependence in daily life activities among individuals with experienced stroke. Despite the fact that this scale is weighted more toward physical disability, it captures (indirectly) other attributes essential to daily activity, including wellbeing, socialization, mood, and cognitive status. The estimate of mRS ranges from 0 (no disability at all) to 6 (dead). The reliability of this scale, including inter-rater and test–retest, lies within moderate and strong limits, respectively (Banks and Marotta, 2007). This scale shortage is associated with its low stroke specificity because it automatically includes other disability causes, such as a previous bone fracture (Kasner, 2006).

2.2.2 National institutes of health stroke scale

The NIHSS scale for quantification of stroke-related neurological impairment consists of 11 neurological examination categories, scored from 0 to 4, with a total score from 0 to 42 (Spilker et al., 1997). On this scale, a higher score indicates more pronounced neurological impairment. The reliability of this scale lies within reasonable limits (Lyden, 2017). To use this scale, one needs special training to reach sufficient reliability and validity. Another shortage of this scale is its inappropriateness for self-report usage or by telephone (Kasner, 2006).

2.2.3 Mini-mental state exam

MMSE is a screening tool used to evaluate cognitive functioning including its five domains (orientation, memory, attention, recollection, and language), with a score rating from 0 to 30 (Folstein et al., 1975). This scale is characterized by better sensitivity for capturing moderate and higher cognitive impairment than mild cognitive impairment (Tombaugh and McIntyre, 1992).

2.2.4 Barthel index

BI was created for the assessment of independence in activities of daily living (Mahoney and Barthel, 1965). This instrument is composed of 10 items, with four possible choices scored as 0, 5, 10, or 15. The possible scores range from 0 to 100. Higher values indicate better functional independence. The reliability of this scale was 0.98 when assessed with Cronbach's α (Shinar et al., 1987). The limitation of this scale is its "ceiling effect" because it does not include many aspects that are important for daily activity, such as emotional disturbances, cognition, and language among others (Kasner, 2006).

2.2.5 EQ-5D-5L

The EQ-5D-5L instrument is composed of two parts, including a descriptive part made up of five different health dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) with five possible options and a thermometer-like visual analog scale (EQ-VAS) numbered from 0 ("worst" HRQL) to 100 ("best" HRQL) to measure overall health (Herdman et al., 2011). The self-described descriptive part of the EQ-5D-5L can be expressed as one of 3,125 different health states, from "No" ("best" HRQL or level 1) to "Extreme" ("worst" HRQL or level 5) problems in all dimensions, or expressed as one index value (EQ index), ranging from slightly <0–1, with higher values indicating a better HRQL. The EQ index mirrors how positive or negative the health state is, depending on the preferences of the country of study. The EQ-VAS is scored by the respondent marking "X" on the scale and separately clarifying the marked point with a number indicating their current health. The self-filled paper version of the EQ-5D-5L in the Lithuanian language has been available since 2014. The self-complete version of the Lithuanian EQ-5D-5L paper was used with formal consent from EuroQol Group with the assigned number 53563. As there is no calculated country-specific EQ index value set for the Lithuanian population, the set from the closest available country is selected, which is from the German population (Ludwig et al., 2018). In the current study, Cronbach's alpha for the measurement was 0.81, while McDonald's Omega was 0.83.

2.2.6 Hospital anxiety and depression scale

The HADS, a self-report screening scale, is composed of two parts, assigned to depression and anxiety severity assessment, respectively (Zigmond and Snaith, 1983). Each part has seven items with four possible options ranging from 0 to 3 according to the psychological distress experienced during the past week, thus generating a score from 0 to 21, with a higher score indicating more pronounced psychological distress. According to previous studies,

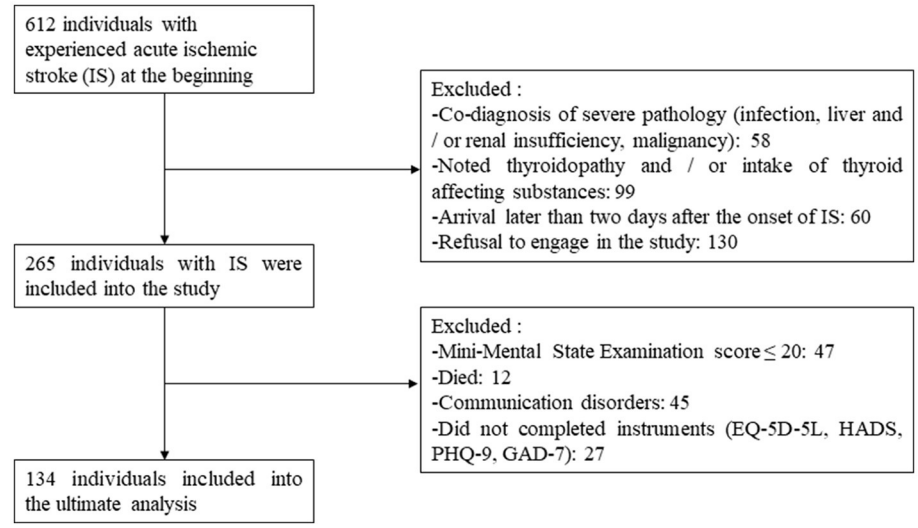


FIGURE 1
Recruitment process.

Cronbach's α of the anxiety part varied from 0.68 to 0.83, while the depression part ranged from 0.67 to 0.90 (Bjelland et al., 2002). The shortcoming of HADS is its dependence on self-reporting which could be impaired because of language and emotional disturbances. A validated Lithuanian version of this instrument (Bunevicius, 1991) was used with permission from the "GL Education Group".

2.2.7 Patient health questionnaire-9

The PHQ-9, a self-report questionnaire for estimating the severity of depression, is composed of nine questions, each of them reflecting depression symptoms in the past 2 weeks, rated from 0 to 3 (Kroenke et al., 2001). The total score can range from 0 to 27, with higher scores reflecting more severe depression. At the end of the questionnaire, there is an additional optional question for global functional impairment assessment. Recently, this scale was validated in a Lithuanian student sample and individuals with anxiety and mood disorders with an estimated reliability (Cronbach's α) of 0.86 (Pranckeviciene et al., 2022; Stanyte et al., 2023). Currently, the Lithuanian version is available on the screener's website (https://www.phqscreeners.com/select-screener). The limitation of this scale is its dependence on intact respondents' communication.

2.2.8 Generalized anxiety disorder-7

The GAD-7 questionnaire was developed for generalized anxiety screening and assessment of its severity (Spitzer et al., 2006). This instrument is composed of seven questions, reflecting anxiety symptoms during the past 2 weeks, with four possible answers ranging from "not at all" to "nearly every day", scored 0 and 3, respectively. Thus, the overall GAD-7 score can range from 0 to 21, with a higher score showing more pronounced symptoms of anxiety. Recently, in Lithuania, the GAD-7 was validated as a first-line anxiety screening tool (Pranckeviciene et al., 2022; Stanyte et al., 2023). For this instrument, Cronbach's α was 0.91. Lithuanian form of instrument is available on the website (https://www.phqscreeners.com/select-screener). The limitation of this scale is its dependence on intact respondents' communication.

2.3 Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics for Windows (version 28) (SPSS Inc, Chicago, IL, USA) and IBM SPSS AMOS 28 (IBM Corp., Armonk, NY, USA). Quantitative data were expressed as the mean (±standard deviation, SD) or median (interquartile range, IQR), with normality checked using the Kolmogorov–Smirnov test. Qualitative data were expressed in number (%).

The reliability of the used questionnaires is expressed as Cronbach's α and McDonald's omega (Hayes and Coutts, 2020). Cronbach's α coefficient estimates between 0.70 and 0.95 were considered to be acceptable (Tavakol and Dennick, 2011). Both the ceiling and floor effects of the EQ-5D-5L health profile, with scores of level "1" or "5" in all dimensions, EQ index, EQ-VAS, HADS for depression and anxiety, HADS total, PHQ-9, and GAD-7 scores, were reported as the proportion of individuals reporting the highest and lowest possible estimates, respectively. A questionnaire was considered to show a ceiling or floor effect if at least 15% of respondents scored the highest or lowest achievable score (Terwee et al., 2007).

The convergent evidence for the EQ-5D-5L, including the separate dimensions of this scale, the EQ index and the EQ-VAS, with other used self-reported questionnaires (BI and NIHSS), was evaluated using Spearman's correlation coefficient. The closeness of co-variation was defined according to the value of the correlation coefficient: ≤0.30 as negligible, 0.31–0.50 as low, 0.51–0.70 as moderate, 0.71–0.90 as high, and 0.91–1.00 as very high (Mukaka, 2012).

Confirmatory factor analysis was performed for validation of the factor structure considered for one (Bilbao et al., 2022) and

TABLE 1 Characteristics of all study patients.

	Total group
Sample size	134
Demographics	
Age, years median (IQR)	66.0 (58.8–73.0)
Age, years mean (SD)	67 (9.6)
Sex, M, *n* (%)	83 (61.9)
Sex, F, *n* (%)	51 (38.1)
Body mass index, median (IQR)	27.7 (24.8–31.8)
Premorbid disability, *n* (%)	6 (3.8)
Used antithrombotic drugs, *n* (%)	50 (37.3)
Chemical thrombolysis, *n* (%)	40 (32.8)
Vascular risk factors	
Arterial hypertension, *n* (%)	100 (74.6)
Atrial fibrillation, *n* (%)	42 (31.3)
Smoking, *n* (%)	32 (23.9)
Diabetes mellitus, *n* (%)	20 (14.9)
Previous cerebral ischemic event, *n* (%)	23 (17.2)
Previous myocardial infarction, *n* (%)	12 (9.0)

F, female; M, male; IQR, interquartile range; NIHSS, National Institutes of Health Stroke Scale.

two factors (Santiago et al., 2021). Analysis of Moment Structures (AMOS) 27.0 software was used to test the model of the EQ-5D-5L using CFA. The proposed thresholds for the CFA fit indices were: CFI > 0.90 adequate and >0.95 good; TLI > 0.90 adequate and >0.95 good; NFI > 0.90 adequate and >0.95 good; RMSEA < 0.08; and χ^2/df with the desired range of 2–5 (Hooper et al., 2008; Brown, 2015). In addition, standardized coefficients for each EQ-5D-5L item were calculated.

The dimensions of the EQ-5D-5L in stroke patients at baseline and after a year were compared using the Wilcoxon signed-rank test. Changes were interpreted according to the Pareto Classification of Health Change (Devlin et al., 2010).

3 Results

Table 1 shows the basic characteristics of study participants. Table 2 shows the main identified characteristics of the used scales. Estimates of the reliability coefficient were within acceptable limits for all scales, except for the HADS depression scale, for which this was marginal (Cronbach's $\alpha = 0.699$). The ceiling effect of the EQ-5D-5L health profile with a full health state of "11111" was highlighted at a significant level in 17.2% of all respondents. In contrast, no floor effect was detected (health state of "55555"). Regarding the EQ index, alongside the same ceiling estimate for the EQ-5D-5L health profile, the floor effect was observed in 0.7% of all respondents. Ceiling and floor effects of the EQ-VAS were observed in 0.7 and 1.5% of individuals, respectively. Of the other scales for the evaluation of psychological distress, only the GAD-7 showed a significant ceiling effect, with a fixed estimate of 41.7%. A significant floor effect was observed for the BI (23.5%).

The convergent validity of the EQ-5D-5L was analyzed, and its correlation with other variables is presented in Table 3. A positive but low correlation was established between the EQ-5D-5L mobility dimension and HADS depression ($r = 0.337$, $p < 0.001$), HADS total ($r = 0.328$, $p < 0.001$), GAD-7 ($r = 0.300$, $p = 0.006$), and NIHSS ($r = 0.413$, $p < 0.001$); between the EQ-5D-5L self-care dimension and NIHSS ($r = 0.483$, $p < 0.001$); between the EQ-5D-5L usual activity dimension and HADS total ($r = 0.306$, $p < 0.001$) and NIHSS ($r = 0.472$, $p < 0.001$); between the EQ-5D-5L pain/discomfort dimension and PHQ-9 ($r = 0.410$, $p < 0.001$) and GAD-7 ($r = 0.312$, $p = 0.004$); and between the EQ-5D-5L anxiety/depression dimension and HADS depression ($r = 0.393$, $p < 0.001$), HADS anxiety ($r = 0.438$, $p < 0.001$), HADS total ($r = 0.495$, $p < 0.001$), PHQ-9 ($r = 0.338$, $p < 0.001$), and GAD-7 ($r = 0.338$, $p < 0.001$). A statistically significant ($p < 0.001$), moderate, negative correlation was found between the EQ-5D-5L mobility ($r = -0.695$, $p < 0.001$) and usual activity dimensions ($r = -0.663$, $p < 0.001$), and a high negative correlation was found between the self-care dimension ($r = -0.756$, $p < 0.001$) and the BI. The correlation between the EQ index and HADS depression ($r = -0.405$, $p < 0.001$), HADS anxiety ($r = -0.339$, $p < 0.001$), HADS total ($r = -0.443$, $p < 0.001$), PHQ-9 ($r = -0.411$, $p < 0.001$), GAD-7 ($r = -0.392$, $p < 0.001$), and NIHSS ($r = -0.371$, $p < 0.001$) was low and negative but positive and moderate with BI ($r = 0.612$, $p < 0.001$). The correlations between other variables were at a negligible correlation level and/or statistically insignificant.

Table 4 shows that the fit of the unidimensional structure was mixed since RMSEA (>0.08) had unacceptable values. On the other hand, the fit of the two-dimensional structure was excellent since both CFI (>0.95) and RMSEA (<0.08) had good values. The two-factor model showed an acceptable fit (RMSEA = 0.045, 90% CI = 0.000–0.145; CFI = 0.996; NFI = 0.983; TLI = 0.991; $\chi^2/df = 1.27$).

The results supporting convergent evidence between isolated factors from the EQ-5D-5L and other applied measures are presented in Table 5, expressed as correlations. Factor 1 (physical) was positively and significantly (0.201–0.377, $p < 0.05$) correlated with the HADS depression, HADS total, PHQ-9, and NIHSS within low correlation limits but negatively (-0.708, $p < 0.001$) and highly correlated with BI. Additionally, a positive, low correlation (0.362–0.478, $p < 0.001$) was established between factor 2 (emotional) and HADS anxiety, HADS depression, PHQ-9, and GAD-7, and a high correlation was established with HADS total ($p < 0.001$). Standardized coefficients for EQ-5D-5L items ranged from 0.55 (anxiety/depression), 0.62 (pain discomfort), 0.82 (mobility), and 0.87 (activities) to 0.92 (self-care).

After 1 year of IS, the EQ-5D-5L data were available for 117 of the included individuals. The comparison of EQ-5D-5L dimensions between two different time points is shown in Table 6. Significant changes were observed in the mobility ($p = 0.013$) and anxiety/depression ($p < 0.001$) dimensions.

4 Discussion

Our results indicate that the descriptive EQ-5D-5L system could be used as a reliable and valid tool for HRQL assessment in individuals living in Lithuania during their hospitalization period

10.3389/fpsyg.2023.1284859

TABLE 2 Characteristics of the scales used in the study population ($n = 134$).

Measures	No. of items	Mean ± SD	Median (IQR)	Min	Max	Ceiling, *n* (%)	Floor, *n* (%)	Cronbach's α
Quality of life								
EQ-5D-5L	5							
Mobility		2.55 ± 1.47	2 (1–4)	1	5	45 (33.6)	21 (15.7)	
Self-care		2.14 ± 1.43	1 (1–3)	1	5	69 (51.5)	14 (10.4)	
Usual activities		2.51 ± 1.47	2 (1–4)	1	5	48 (35.8)	20 (14.9)	
Pain/discomfort		1.99 ± 1.12	2 (1–3)	1	5	62 (46.3)	4 (3.0)	
Anxiety/depression		1.73 ± 1.02	1 (1–2)	1	5	75 (56.0)	4 (0.3)	
EQ-5D-5L total		10.92 ± 4.94	10 (7–15)	5	25	23 (17.2)	0 (0.0)	0.809
EQ index		0.69 ± 0.32	0.82 (0.47–0.93)	−0.34	1	23 (17.2)	1 (0.7)	
EQ VAS		58.36 ± 23.81	60 (50–80)	0	100	1 (0.7)	2 (1.5)	
Psychological distress								
HADS								
HADS depression	7	4.87 ± 3.70	4 (2–7)	0	16	7 (5.2)	1 (0.7)	0.699
HADS anxiety	7	4.79 ± 3.70	4 (2–7)	0	18	12 (8.9)	1 (0.7)	0.751
HADS total	14	9.66 ± 6.14	8 (6–14)	0	33	2 (1.5)	1 (0.7)	0.781
PHQ-9	9	5.18 ± 4.89	4 (2–7)	0	21	14 (11.2)	3 (2.4)	0.796
GAD-7	7	2.95 ± 3.44	2 (0–5)	0	14	35 (41.7)	3 (3.6)	0.826
Functional independence								
Barthel index	10	69.4 ± 32.79	80 (50–95)	0	100	48 (25.4)	13 (6.9)	0.949
Neurological impairment								
NIHSS	11	8.94 ± 6.98	7 (4–12)	0	39	6 (2.9)	1 (0.5)	0.805

VAS, visual analog scale; HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient Health Questionnaire 9; GAD-7, Generalized Anxiety Disorder assessment-7; NIHSS, National Institutes of Health Stroke Scale; SD, standard deviation.

due to IS. Furthermore, the presented results suggest the presence of two EQ-5D-5L factors in individuals who have experienced IS.

This study establishes that the EQ-5D-5L health profile shows no floor effect but has a significant ceiling effect. This is consistent with the ceiling effect described in the post-stroke population in Taiwan, which was even higher (20%) (Chen et al., 2016). Another study, which included native Polish speakers, found a much lower ceiling effect of 5.6% (Golicki et al., 2015). In general, it is agreed that the EQ-5D-5L is prone to a large ceiling effect because of its nature in measuring more aspects of negative health than positive health (Feng et al., 2021). In addition, the tendency for more positive HRQL self-evaluation in Lithuania may be associated with cultural and historical (post-Soviet) aspects, such as denial of psychological distress (Gailiene, 2021). In terms of the EQ index and EQ-VAS, the ceiling and floor effects were non-significant.

The convergent validity of the EQ-5D-5L health profile justifies the identified significant correlations between the anxiety/depression dimension and all other used scales for measuring psychological distress, as well as between the pain/discomfort dimension and PHQ-9 and GAD-7. EQ-5D-5L dimensions such as mobility, self-care, and usual activities correlated more with scales that included a mobility component, namely, the NIHSS scale, and even with BI. In addition, the latter EQ-5D-5L dimensions were correlated with the HADS total, and the mobility dimension was correlated with HADS depression and GAD-7. The EQ index showed a significant correlation with all included instrument scores, adding additional justification for the convergent validity of the descriptive EQ-5D-5L system. As for EQ-VAS, no significant correlations point to unjustified convergent validity of this EQ-5D-5L component. An established difference could be attributed to the EQ-5D-5L health profile and the EQ index to social perspectives and EQ-VAS to personal perspectives. Furthermore, another explanations could be that the EQ-VAS is a wider construct than the EQ-5D-5L health profile; misinterpretation of the EQ-VAS filling instructions; and difficulty in understanding this two-pole scale (Feng et al., 2014), especially keeping in mind that our study population consisted of individuals with an organically injured brain—the substrate for cognition. In addition, the study from Taiwan did not show EQ-VAS power for predicting rehabilitation outcomes after stroke (Kainz et al., 2021).

Our study revealed the existence of two factors of the EQ-5D-5L, which is in line with a study exploring the validity of this instrument among individuals with heart disease (Gao et al., 2019). The latter study separated only the EQ-5D-5L anxiety/depression dimension into the second factor, while our results additionally identified the pain/discomfort dimension. In our study, we highlighted that the first factor, composed of EQ-5D-5L mobility, self-care, and usual activity dimensions, could

TABLE 3 Convergent evidence of the EQ-5D-5L with HADS, PHQ-9, GAD-7, NIHSS, and Barthel Index in the overall sample ($n = 134$).

Scales	EQ-5D-5L						
	Mobility	Self-care	Usual activities	Pain/discomfort	Anxiety/depression	EQ index	EQ VAS score
HADS							
HADS depression	0.337**	0.228*	0.249*	0.220*	0.393**	−0.405**	−0.194
HADS anxiety	0.216*	0.172*	0.264*	0.226*	0.438**	−0.339**	−0.159
HADS total	0.328**	0.237*	0.306**	0.266*	0.495**	−0.443**	−0.210*
PHQ-9	0.270*	0.255*	0.244*	0.410**	0.338**	−0.411**	−0.144
GAD-7	0.300*	0.160	0.284*	0.312*	0.350*	−0.392*	−0.202
NIHSS	0.413**	0.483**	0.472**	0.006	0.112	−0.371**	−0.071
Barthel Index	−0.695**	−0.756**	−0.663**	−0.139	−0.112	0.612**	0.209*

VAS, a visual analog scale; HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient health questionnaire-9; GAD-7, Generalized anxiety disorder assessment-7; NIHSS, National Institutes of Health Stroke Scale.
*$p < 0.05$.
**$p < 0.001$.

TABLE 4 Confirmatory factor analysis of two measurement models of EQ-5D-5L.

	χ^2/df	CFI	TLI	NFI	RMSEA (90% CI)
1-factor model	2.87	0.968	0.936	0.953	0.119 (0.049–0.193)
2-factor model	1.27	0.996	0.936	0.983	0.045 (0.000–0.145)

CFI, comparative fit index; TLI, Tucker–Lewis Index; NFI, non-normal fit index; RMSEA, root-mean-square error of approximation; 90% CI, 90% confidence interval of the RMSEA.

TABLE 5 Convergent evidence of the factors of EQ-5D-5L with HADS, PHQ-9, GAD-7, NIHSS, and Barthel Index in the overall sample.

	EQ-5D-5L	
	Factor 1	Factor 2
HADS		
HADS depression	0.371**	0.362**
HADS anxiety	0.201*	0.478**
HADS total	0.350**	0.505**
PHQ-9	0.301**	0.449**
GAD-7	0.271*	0.451**
NIHSS	0.377**	−0.016*
Barthel Index	−0.708**	−0.098*

HADS, Hospital Anxiety and Depression Scale; PHQ-9, Patient health questionnaire-9; GAD-7, Generalized anxiety disorder assessment-7; NIHSS, National Institutes of Health Stroke Scale.
*$p < 0.05$.
**$p < 0.001$.

be attributed to the physical component of this scale, while the other two dimensions—pain/discomfort and anxiety/depression—could be attributed to the emotional one. According to our results, research from Australia with indigenous people found identical EQ-5D-5L two-dimensional latent factor composition (Santiago et al., 2021). On the other hand, in Spain, evaluating psychometrics of this instrument in individuals with depression showed uni-dimensionality of latent factors (Bilbao et al., 2022). Different results could be attributed to different clinical entities (stroke, strictly organic brain disease with physical and emotional consequences, vs. depression, with a more pronounced emotional component).

Convergent analysis of revealed factors substantiated their relevance, with a significant correlation between the first factor and NIHSS and BI and between the second factor and applied questionnaires dedicated to psychological distress assessment (HADS, PHQ-9, and GAD-7). Here, the low positive correlation between NIHSS and the first factor could be attributed to differences in the evaluation of NIHSS and EQ-5D-5L in time and less sensitivity of the latter measure to neurologic deficits evaluated with NIHSS such as neglect and visual disturbances (van der Ende et al., 2023).

Finally, our results showed that HRQL was not static after stroke. An unadjusted analysis of the EQ-5D-5L health profile confirmed meaningful changes in responses to the mobility and anxiety/depression dimensions. Here, mobility improved, but anxiety/depression deteriorated.

5 Strengths, limitations, and applications

The strengths of the present study are the participation of three different centers, a large enough sample size, and the use of validated scales. The limitations of the study were the exclusion of individuals with communication disorders, the unavailability of radiological data related to stroke volume and place, and the lack of comparisons made with other HRQL instruments. This study further expands the territory of usage of the EQ-5D-5L instrument for HRQL assessment in individuals after IS, adding the country

TABLE 6 Descriptive statistics of the EQ-5D-5L dimensions in stroke patients at baseline and after a year.

Dimension	Baseline, median (IQR)	Follow-up, median (IQR)	*p*
	N = 117		
Mobility	2.55 ± 1.47	2.26 ± 1.34	**0.013**
Self-care	2.11 ± 1.40	2.08± 1.43	0.795
Usual activities	2.49 ± 1.45	2.43 ± 1.53	0.695
Pain/discomfort	1.97 ± 1.11	2.07 ± 1.19	0.443
Anxiety/depression	1.72 ± 1.01	2.19 ± 1.30	**<0.001**

Bolded, $p < 0.05$.

of Lithuania. This validated instrument creates an opportunity for further clinical and economic research dedicated to improving IS-associated HRQL in Lithuania.

6 Conclusion

This study adds knowledge of the psychometric properties of the EQ-5D-5L instrument in individuals who have experienced IS in Lithuania. The research confirmed that the EQ-5D-5L instrument and its derivative EQ index are a valid and reliable tool for HRQL assessment in individuals at the end of the first week after IS. In addition, the analysis revealed two factors behind the EQ-5D-5L health profile, with possible physical and emotional dimensions. The data did not support the validity of overall health expressed as EQ-VAS scoring in these individuals. Our study supports further research using the EQ-5D-5L instrument for HRQL assessment in individuals who have experienced stroke.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Regional Biomedical Research Ethics Committee (Permission Numbers: BE-2-11/2013; P2-BE-2-11/2013). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ST: Investigation, Methodology, Writing—original draft. JB: Investigation, Methodology, Writing—review & editing. AP: Methodology, Writing—review & editing. NK: Project administration, Writing—review & editing. NM: Conceptualization, Funding acquisition, Resources, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by a Grant (No. S-MIP-23-103) from the Research Council of Lithuania.

Acknowledgments

The authors would like to thank the patients who participated in the study.

Conflict of interest

JB, is a consultant at Cronos and Saulius Taroza, and reports personal fees from Berlin Chemie Menarini Baltic.

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.

Publisher's note

References

Banks, J. L., and Marotta, C. A. (2007). Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. *Stroke* 38, 1091–1096. doi: 10.1161/01.STR.0000258355.23810.c6

Bilbao, A., Martín-Fernández, J., García-Pérez, L., Mendezona, J. I., Arrasate, M., Candela, R., et al. (2022). Psychometric properties of the EQ-5D-5L in patients with major depression: factor analysis and Rasch analysis. *J. Ment. Health* 31, 506–516. doi: 10.1080/09638237.2021.1875422

Bjelland, I., Dahl, A. A., Haug, T. T., and Neckelmann, D. (2002). The validity of the hospital anxiety and depression scale. An updated literature review. *J. Psychosom. Res.* 52, 69–77. doi: 10.1016/S0022-3999(01)00296-3

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., and Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front. Public Health* 6, 149. doi: 10.3389/fpubh.2018.00149

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Publications.

Bunevicius, A., Donovan, L., and Sheehan, J. (2022). Health related quality of life trajectories after stereotactic radiosurgery for brain metastases: a systematic review. *J. Neurooncol.* 159, 319–331. doi: 10.1007/s11060-022-04067-8

Bunevicius, R. Z. S. (1991). Correlations between MMPI and HAD scale. *Psychol. Res. Arch. Lithuan. Univ.* 11, 95–102. doi: 10.15388/Psichol.1991.11.9062

Burkauskas, J., Mickuviene, N., Brozaitiene, J., Staniute, M., Podlipskyte, A., Rastenyte, D., et al. (2014). Gene-environment interactions connecting low triiodothyronine syndrome and outcomes of cardiovascular disease (GET-VASC): study protocol. *Biol. Psychiatry Psychopharmacol.* 16, 66–73.

Cadilhac, D. A., Dewey, H. M., Vos, T., Carter, R., and Thrift, A. G. (2010). The health loss from ischemic stroke and intracerebral hemorrhage: evidence from the North East Melbourne Stroke Incidence Study (NEMESIS). *Health Qual Life Outcomes* 8, 49. doi: 10.1186/1477-7525-8-49

Cameron, L. J., and Wales, K. (2022). Self-reported quality of life following stroke: a systematic review of instruments with a focus on their psychometric properties. *Qual. Life Res.* 31, 329–342. doi: 10.1007/s11136-021-02944-9

Campbell, B. C. V., and Khatri, P. (2020). Stroke. *Lancet* 396, 129–142. doi: 10.1016/S0140-6736(20)31179-X

Chen, G., and Olsen, J. A. (2020). Filling the psycho-social gap in the EQ-5D: the empirical support for four bolt-on dimensions. *Qual. Life Res.* 29, 3119–3129. doi: 10.1007/s11136-020-02576-5

Chen, P., Lin, K.-C., Liing, R.-J., Wu, C.-Y., Chen, C.-L., and Chang, K.-C. (2016). Validity, responsiveness, and minimal clinically important difference of EQ-5D-5L in stroke patients undergoing rehabilitation. *Qual Life Res.* 25, 1585–1596. doi: 10.1007/s11136-015-1196-z

Chen, Q., Cao, C., Gong, L., and Zhang, Y. (2019). Health related quality of life in stroke patients and risk factors associated with patients for return to work. *Medicine* 98, e15130. doi: 10.1097/MD.0000000000015130

Coons, S. J., Rao, S., Keininger, D. L., and Hays, R. D. (2000). A comparative review of generic quality-of-life instruments. *Pharmacoeconomics* 17, 13–35. doi: 10.2165/00019053-200017010-00002

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality prediction with a single general self-rated health question. A meta-analysis. *J. Gen. Intern. Med.* 21, 267–275. doi: 10.1111/j.1525-1497.2005.00291.x

Devlin, N. J., Parkin, D., and Browne, J. (2010). Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. *Health Econ.* 19, 886–905. doi: 10.1002/hec.1608

Dhamoon, M. S., Moon, Y. P., Paik, M. C., Boden-Albala, B., Rundek, T., Sacco, R. L., et al. (2010). Quality of life declines after first ischemic stroke. The Northern Manhattan Study. *Neurology* 75, 328–334. doi: 10.1212/WNL.0b013e3181ea9f03

Feng, Y., Parkin, D., and Devlin, N. J. (2014). Assessing the performance of the EQ-VAS in the NHS PROMs programme. *Qual. Life Res.* 23, 977–989. doi: 10.1007/s11136-013-0537-z

Feng, Y. S., Kohlmann, T., and Janssen, M. F. (2021). Psychometric properties of the EQ-5D-5L: a systematic review of the literature. *Qual. Life Res.* 30, 647–673. doi: 10.1007/s11136-020-02688-y

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Gailiene, D. (2021). *Ka jie mums padare: Lietuvos gyvenimas traumu psichologijos Žvilgsniu*. Vilnius: Tyto alba, 246.

Gall, C., Franke, G. H., and Sabel, B. A. (2010). Vision-related quality of life in first stroke patients with homonymous visual field defects. *Health Qual. Life Outcomes* 8, 33. doi: 10.1186/1477-7525-8-33

Gao, L., Moodie, M., and Chen, G. (2019). Measuring subjective wellbeing in patients with heart disease: relationship and comparison between health-related quality of life instruments. *Qual. Life Res.* 28, 1017–1028. doi: 10.1007/s11136-018-2094-y

Golicki, D., Niewada, M., Buczek, J., Karlińska, A., Kobayashi, A., Janssen, M. F., et al. (2015). Validity of EQ-5D-5L in stroke. *Qual. Life Res.* 24, 845–850. doi: 10.1007/s11136-014-0834-1

Hatano, S. (1976). Experience from a multicentre stroke register: a preliminary report. *Bull. World Health Organ.* 54, 541–553.

Hayes, A. F., and Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *But Commun. Methods Meas.* 14, 1–24. doi: 10.1080/19312458.2020.1718629

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual. Life Res.* 20, 1727–1736. doi: 10.1007/s11136-011-9903-x

Hooper, D., Coughlan, J., and ve Mullen, M. R. (2008). Structural equation modelling: guidelines for determining model fit. *Electr. J. Bus. Res. Methods*. 6, 53–60.

Kainz, A., Meisinger, C., Linseisen, J., Kirchberger, I., Zickler, P., Naumann, M., et al. (2021). Changes of health-related quality of life within the 1st year after stroke-results from a prospective stroke cohort study. *Front. Neurol.* 12, 715313. doi: 10.3389/fneur.2021.715313

Kaplan, R. M. (1990). Behavior as the central outcome in health care. *Am. Psychol.* 45, 1211–1220. doi: 10.1037/0003-066X.45.11.1211

Kaplan, R. M. (2003). The significance of quality of life in health care. *Qual. Life Res.* 12(Suppl. 1), 3–16. doi: 10.1023/A:1023547632545

Karimi, M., and Brazier, J. (2016). Health, health-related quality of life, and quality of life: what is the difference? *Pharmacoeconomics* 34, 645–649. doi: 10.1007/s40273-016-0389-9

Kasner, S. E. (2006). Clinical interpretation and use of stroke scales. *Lancet Neurol.* 5, 603–612. doi: 10.1016/S1474-4422(06)70495-1

Kranciukaite-Butylkiniene, D. (2014). *Quality of Life After Stroke: The EROS Study in Urban Lithuania*. Göteborg: Nordic School of Public Health NHV.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Lau, V. I., Johnson, J. A., Bagshaw, S. M., Rewa, O. G., Basmaji, J., Lewis, K. A., et al. (2022). Health-related quality-of-life and health-utility reporting in critical care. *World J. Crit. Care Med.* 11, 236–245. doi: 10.5492/wjccm.v11.i4.236

Ludwig, K., Graf von der Schulenburg, J. M., and Greiner, W. (2018). German value set for the EQ-5D-5L. *Pharmacoeconomics* 36, 663–674. doi: 10.1007/s40273-018-0615-8

Lyden, P. (2017). Using the National Institutes of Health Stroke Scale: a cautionary tale. *Stroke* 48, 513–519. doi: 10.1161/STROKEAHA.116.015434

Mahoney, F. I., and Barthel, D. W. (1965). Functional evaluation: the barthel index. *Md State Med. J.* 14, 61–65. doi: 10.1037/t02366-000

Mar, J., Masjuan, J., Oliva-Moreno, J., Gonzalez-Rojas, N., Becerra, V., Casado, M. Á., et al. (2015). Outcomes measured by mortality rates, quality of life and degree of autonomy in the first year in stroke units in Spain. *Health Qual. Life Outcomes* 13, 36. doi: 10.1186/s12955-015-0230-8

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71.

Pequeno, N. P. F., Cabral, N. L. D. A., Marchioni, D. M., Lima, S. C. V. C., and Lyra, C. D. O. (2020). Quality of life assessment instruments for adults: a systematic review of population-based studies. *Health Qual. Life Outcomes* 18, 208. doi: 10.1186/s12955-020-01347-7

Pranckeviciene, A., Saudargiene, A., Gecaite-Stonciene, J., Liaugaudaite, V., Griskova-Bulanova, I., Simkute, D., et al. (2022). Validation of the patient health questionnaire-9 and the generalized anxiety disorder-7 in Lithuanian student sample. *PLoS ONE* 17, e0263027. doi: 10.1371/journal.pone.0263027

Revicki, D. A., Kleinman, L., and Cella, D. (2014). A history of health-related quality of life outcomes in psychiatry. *Dialogues Clin. Neurosci.* 16, 127–135. doi: 10.31887/DCNS.2014.16.2/drevicki

Santiago, P. H. R., Haag, D., Macedo, D. M., Garvey, G., Smith, M., Canfell, K., et al. (2021). Psychometric properties of the EQ-5D-5L for aboriginal Australians: a multi-method study. *Health Qual. Life Outcomes* 19, 81. doi: 10.1186/s12955-021-01718-8

Shinar, D., Gross, C. R., Bronstein, K. S., Licata-Gehr, E. E., Eden, D. T., Cabrera, A. R., et al. (1987). Reliability of the activities of daily living scale and its use in telephone interview. *Arch. Phys. Med. Rehabil.* 68, 723–728.

Spilker, J., Kongable, G., Barch, C., Braimah, J., Bratina, P., Daley, S., et al. (1997). Using the NIH Stroke Scale to assess stroke patients. The NINDS rt-PA stroke study group. *J. Neurosci. Nurs.* 29, 384–392. doi: 10.1097/01376517-199712000-00008

Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch. Intern. Med.* 166, 1092–1097. doi: 10.1001/archinte.166.10.1092

Stanyte, A., Fineberg, N. A., Podlipskyte, A., Gecaite-Stonciene, J., Macijauskiene, J., et al. (2023). Validation of the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder-7 in Lithuanian individuals with anxiety and mood disorders. *J. Psychiatr. Res.* 164, 221–228. doi: 10.1016/j.jpsychires.2023.06.027

Stenman, U., Hakama, M., Knekt, P., Aromaa, A., Teppo, L., Leinonen, J., et al. (2010). Measurement and modeling of health-related quality of life. *Epidem. Demog. Public Health* 195, 130–135.

Tavakol, M., and Dennick, R. (2011). Making sense of Cronbach's alpha. *Int. J. Med. Educ.* 2, 53–55. doi: 10.5116/ijme.4dfb.8dfd

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* 60, 34–42. doi: 10.1016/j.jclinepi.2006.03.012

Tombaugh, T. N., and McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *J. Am. Geriatr. Soc.* 40, 922–935. doi: 10.1111/j.1532-5415.1992.tb01992.x

van der Ende, N. A. M., den Hartog, S. J., Broderick, J. P., Khatri, P., Visser-Meily, J. M. A., van Leeuwen, N., et al. (2023). Disentangling the association between neurologic deficits, patient-reported impairments, and quality of life after ischemic stroke. *Neurology* 100, e1321–e1328. doi: 10.1212/WNL.0000000000206747

Xie, J., Wu, E. Q., Zheng, Z. J., Croft, J. B., Greenlund, K. J., Mensah, G. A., et al. (2006). Impact of stroke on health-related quality of life in the noninstitutionalized population in the United States. *Stroke* 37, 2567–2572. doi: 10.1161/01.STR.0000240506.34616.10

Zigmond, A. S., and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatr. Scand.* 67, 361–370. doi: 10.1111/j.1600-0447.1983.tb09716.x

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 04 January 2024
DOI 10.3389/fpsyg.2023.1270437

OPEN ACCESS

EDITED BY
Alessio Facchin,
University of Milano-Bicocca, Italy

REVIEWED BY
Billino Jutta,
University of Giessen, Germany
Daniela De Bartolo,
Vrije Universiteit Amsterdam, Netherlands

*CORRESPONDENCE
Jennifer Randerath
✉ J_Randerath@hotmail.com

RECEIVED 31 July 2023
ACCEPTED 28 November 2023
PUBLISHED 04 January 2024

CITATION
Stoll SEM, Bauer I, Hopfer K, Lamberty J, Lunz V, Guzmán Bausch F, Höflacher C, Kroliczak G, Kalénine S and Randerath J (2024) Diagnosing homo digitalis: towards a standardized assessment for digital tool competencies.
Front. Psychol. 14:1270437.
doi: 10.3389/fpsyg.2023.1270437

Diagnosing homo digitalis: towards a standardized assessment for digital tool competencies

Sarah E. M. Stoll[1,2,3], Isabel Bauer[1,2], Karen Hopfer[1], Judith Lamberty[1], Verena Lunz[1], Francesca Guzmán Bausch[1], Cosima Höflacher[1], Gregory Kroliczak[4,5], Solène Kalénine[6] and Jennifer Randerath[1,2,7]*

[1]Department of Psychology, University of Konstanz, Konstanz, Germany, [2]Lurija Institute for Rehabilitation Science and Health Research, Kliniken Schmieder, Allensbach, Germany, [3]Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria, [4]Cognitive Neuroscience Center, Action and Cognition Laboratory, Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Poznan, Poland, [5]Department of Clinical Neuropsychology, Nicolaus Copernicus University in Toruń Collegium Medicum, Bydgoszcz, Poland, [6]Sciences Cognitives Et Sciences Affectives, University of Lille, Lille, France, [7]Outpatient Unit for Research, Teaching, and Practice, Faculty of Psychology, University of Vienna, Vienna, Austria

Introduction: In the 21st century, digital devices have become integral to our daily lives. Still, practical assessments designed to evaluate an individual's digital tool competencies are absent. The present study introduces the "Digital Tools Test" ("DIGI"), specifically designed for the evaluation of one's proficiency in handling common applications and functions of smartphones and tablets. The DIGI assessment has been primarily tailored for prospective use among older adults and neurological patients with the latter frequently suffering from so-called apraxia, which potentially also affects the handling of digital tools. Similar to traditional tool use tests that assess tool-selection and tool-action processes, the DIGI assessment evaluates an individual's ability to select an appropriate application for a given task (e.g., creating a new contact), their capacity to navigate within the chosen application and their competence in executing precise and accurate movements, such as swiping.

Methods: We tested the implementation of the DIGI in a group of 16 healthy adults aged 18 to 28 years and 16 healthy adults aged 60 to 74 years. All participants were able to withstand the assessment and reported good acceptance.

Results: The results revealed a significant performance disparity, with older adults displaying notably lower proficiency in the DIGI. The DIGI performance of older adults exhibited a correlation with their ability to employ a set of novel mechanical tools, but not with their ability to handle a set of familiar common tools. There was no such correlation for the younger group.

Conclusion: In conclusion, this study introduces an innovative assessment tool aimed at evaluating common digital tool competencies. Our preliminary results demonstrate good acceptance and reveal expected group differences. For current cohorts of older adults, the results seem to indicate that the ability to use novel tools may aid digital tool use. In the next step, the psychometric properties of the DIGI assessment should be evaluated in larger and more diverse samples. The advancement of digital tool competency assessments and rehabilitation strategies is essential when we aim at facilitating societal inclusion and participation for individuals in affected populations.

KEYWORDS

digital tools, aging, digital competencies, assessment, neurorehabilitation, inclusion, digital literacy, novel tools

Introduction

In daily life and society, Information Communication Technologies like the internet, smartphones, tablets, and applications (apps) have become ubiquitous. Proficiency in these technologies and broad digital competencies are important assets for participation in the working world (Oberländer et al., 2020). The concept of "digital competence" was recognized as one of the eight core competencies for lifelong learning by the European Parliament and Council as early as 2006 (European Parliament, 2006). Its significance extends beyond the professional world since the activities of daily living are increasingly shaped by digitalization. In numerous aspects of our lives, digital technologies have emerged as the most convenient means of access. For example, in the realm of transportation and travel, we simply call an Uber via an app or find the nearest subway station with the "maps" application on our smartphones. These digital approaches offer advantages, including flexibility and mobility (Quamar et al., 2020). Arguably, one of the most pivotal roles played by modern digital technologies is in the domain of communication. Instant messaging (e.g., WhatsApp, Messenger), email services, social networking platforms (e.g., Instagram, Facebook) and video conferencing (e.g., Skype, Zoom) nowadays are common (Quamar et al., 2020). Typically, these communication tools are accessed through the use of smartphones.

The role of participation in digital opportunities is particularly evident across different demographic groups. Among the younger population, aged between 20 and 25, digital tools have emerged as the primary medium for communication. In fact, owning a smartphone is considered by this age group as an almost indispensable component of social interaction, and those without such a device are perceived to be partially excluded from these interactions (Möller, 2016). A longitudinal study in a Finnish sample showed that also in middle-aged and older persons, the perceived necessity to own and use information and communication technology (such as smartphones and tablets) was growing (Wilska and Kuoppamäki, 2017). In the case of older adults and individuals with medical conditions, especially in the context of eHealth and mHealth (i.e., the provision of healthcare services through Information Communication Technologies, particularly smart mobile devices), the significance of these technologies has been steadily growing. As preventive measures or complements to traditional medical care, mobile health apps are becoming increasingly accessible via smartphones or tablets. Unfortunately, the adoption of smart mobile devices is still less prevalent in older age groups, even though older adults may benefit the most from telemedical apps and mHealth communication (Chiarini et al., 2013; Li et al., 2014; Changizi and Kaveh, 2017). To close these gaps there is a need to analyze potential factors contributing to non-use.

One important factor for non-use or inappropriate use presents an inadequate understanding of how to properly operate these devices. There are several potential challenges older adults might face when attempting to navigate digital devices. First, there appear good news when looking at overall usability of smartphones and tablets. Kortum and Sorber (2015) who investigated usability ratings of the most popular applications on iOS and Android OS among more than 3,000 participants reported high usability ratings. Also in the older population, there is a positive reception of smart mobile technologies: in a recent study, Brunzini et al. (2023) found that older Italian citizens regarded digital devices, including smartphones and tablets, as quite useable and learnable. Moreover, their small pilot sample demonstrated only few errors when operating these devices for social support, and entertainment purposes. However, usability and performance measures frequently seem to dissociate in older adults. For example, in a study comparing touchscreen versus keyboard use in two tasks, Sonderegger et al. (2016) found that while older adults were equally effective at solving text input- and menu selection-tasks as their younger counterparts, they performed less efficient. At the same time the perceived usability of smartphones was rather positive in older adults. Multiple obstacles faced by senior citizens were identified by McGaughey et al. (2013) or Gomez-Hernandez et al. (2023) in their reviews: Some difficulties can be attributed to the device itself, such as the small size of the gadget, others depend on characteristics of the user, such as physical and cognitive limitations or a lack of confidence and training. Furthermore, studies suggest that age, together with educational background, may have an influence on the ability to solve technology-associated problems (Ertl et al., 2020).

However, non-use due to reduced competencies does not merely pertain to healthy older adults, but also to persons with cognitive disabilities, for example after stroke. We propose that digital tool competencies is also a highly relevant topic in the context of neurorehabilitation. Strikingly, limb apraxia, known as a disorder of (traditional) tool use (Goldenberg, 2013; Randerath, 2023), is a frequent consequence of brain damage such as stroke with a prevalence of 28–37% among stroke survivors (Donkervoort et al., 2000). The term "limb apraxia" refers to disorders of learned and purposeful movements (Liepmann, 1900; Heilman and Rothi, 1993). When applying the traditional tool use assessments using the DILA-S in stroke patients (Buchmann and Randerath, 2017; Buchmann et al., 2020a,b), our patients' left us with the impression that next to their common tool competencies (i.e., how to use a fork or a toothbrush), their digital tool competencies (i.e., send a note or picture to their relatives using a messenger-application) are just as important to them for their ADLs (activities of daily living) and participation. From our observations, there are valid concerns surrounding the capacity of stroke patients to navigate digital devices. Lastly, apraxia is only one of a vast variety of potential syndromes and disorders after stroke that may affect digital tool use. Other stroke-associated symptoms affecting motor, perceptual, communicative, or cognitive abilities such as hemiplegia, hemineglect, aphasia, and deficits concerning concentration and memory are potential influencing factors that also may detrimentally impact digital competencies. Another concern regarding the capacity of stroke patients to operate digital devices relates to the advanced age of many individuals in this patient group (Busch and Kuhnert, 2017). Therefore, it is important to first investigate the digital tool use competencies in healthy older adults.

Considering the profound impact of Information Communication Technologies on ADLs, in their review Quamar et al. (2020) conclude that it "marks a paradigm shift in the way we assess and measure everyday functioning". The digitalization drives the need for standardized tests of basic digital skills to be considered for ADL assessments, contributing to the "paradigm shift".

The assessment of individual difficulties in common digital tool competencies seems an important step towards characterizing an

Abbreviations: DIGI, digital tools test; FTT, familiar tools test; NTT, novel tools test.

individual's problem also before offering a tailored training intervention. There are strong efforts to enhance digital accessibility for the older population, e.g., by designing special user interfaces for older adults (Arab et al., 2013; Sakdulyatham et al., 2017) or providing smartphone training classes (Zhao et al., 2020). A standardized assessment could be useful to evaluate the success of such an intervention. Despite the decent amount of tests for general or specific technological knowledge and skills among high school and college students (for an overview see Covello and Lei, 2010), standardized instruments for the assessment of digital competencies in the general population are scarce. Existing assessments rely on self-report questionnaires rather than practical tasks (Ferrari, 2013; Lu et al., 2017; Karnoe et al., 2018; Zhao et al., 2020) or they focus on device usability (Sonderegger et al., 2016; Brunzini et al., 2023).

Inspired by our clinical work, we developed a novel pragmatic assessment for digital tool competencies. The major goal of this manuscript is to introduce the so-called DIGI (DIGItal tools test). This instrument aims to assess fundamental digital tool competencies focusing on elementary tasks associated with the utilization of smart mobile devices, namely smartphones and tablets. It evaluates participants' performance regarding their ability to select an adequate application (selection), successful navigation inside the application (production) and minimize motor-related errors (motor error). The DIGI has been developed especially for use in older adults and neurologic patients. In the present pilot study, we sought to test the feasibility and acceptance of the DIGI in a sample of healthy young and older adults participating in the assessment. The study further involved a comparative analysis between the two cohorts and included a correlational analysis with performance in the traditional novel and familiar tool use tests of the Diagnostic Instrument for Limb Apraxia (DILA-S).

The current study

Despite unprecedented opportunities of smart mobile devices in supporting independence and healthcare for older adults and neurological patients, many older individuals are hesitant to use these devices due to a lack of competence or because brain damage may have impaired their ability to use these tools. A prerequisite for administering adequate digital tool use training is the standardized assessment of abilities and difficulties in handling smart mobile devices. Currently, there is a lack of a suitable assessment tool for this purpose. In the current study, we aim to address this gap by introducing a newly developed assessment for evaluating digital tool use competencies, named DIGI. This assessment evaluates a set of everyday skills and tasks in operating smartphones or tablets, like saving a contact or connecting the device to the power socket for charging. Performance is evaluated based on correctly choosing (selection) and using (production) the essential features to handle each task, as well as on movement-related mistakes (motor error).

We anticipated no drop outs, good acceptability and that the group of older adults will show significantly more difficulties in handling digital devices compared to the younger group with significantly lower selection and production scores and significantly more movement-related mistakes than the younger group. We further explored whether the proficiency to use traditional novel versus familiar tools would correlate with the ability to use modern smart mobile devices in the young as well as in the older group.

Methods

The study was approved by the ethics committee of the University of Konstanz (#15/2020) and conducted in accordance with the declaration of Helsinki. All participants gave informed consent before taking part in the study. Post hoc power-analyses can be found in the Supplementary material, Table 1.

Participants

Data was collected from March 2019 to July 2019. The younger sample consisted of 16 subjects ranging between 18 and 28 years ($M = 23.50$, $SD = 2.68$), with half of them being female. The older adults sample included 16 participants, aged between 60 and 74 years ($M = 64.25$, $SD = 3.99$), with nine of them being female. None of the participants showed signs of cognitive impairment as evidenced by their DemTect Scores ≥ 13 (Kalbe et al., 2004). Two subjects, one in each group, indicated to be left-handed. Hand sensibility, assessed with the two-point discrimination test (for a detailed description see Hunter et al., 1990) did not differ between the older and younger group ($U = 119.5$, $z = -0.34$, $p = 0.752$).

DIGI

For a comprehensive description of the DIGI, please refer to the manual, booklets and evaluation sheets (available at https://kops.uni-konstanz.de/entities/publication/09b43e22-1e78-4561-9833-9eaa7963f38f). The DIGI was developed to assess the skills in handling digital devices. During the assessment, participants are tasked with completing everyday-like assignments using a smart mobile device. The experimenter evaluates the participant's performance using an evaluation sheet, considering the successful selection of an adequate application, the effective navigation inside the application and the skillfulness of the motor movement when interacting with the device. The DIGI assessment consists of two versions, denoted as A and B, which cater to both, smartphone, and tablet, compatible with the operating systems iOS and Android. Booklets and evaluation sheets are available for both operating systems. In the present pilot study, all participants used the Android-based devices. Each of the two versions comprises the same two practice trials (see AB 00.1 and AB 00.2 in Table 1), eight tasks for smartphone and seven tasks for tablet. Parallel-items that were chosen for their close resemblance were: A01-B01; A02-B02; A03-B03; A04-B04; A05-B07; A06-B08; A07-B05; A08-B06 (please note, whether both subsets are solved in a similar manner will be looked at in a subsequent study evaluating psychometric properties by use of a larger sample). Notably, two tasks involving the phone function (A 02 *answer a call*, B 02 *make a call*) are exclusive to the smartphone version. The remaining tasks are identical for smartphone and tablet. Practice trials are excluded from the evaluation, since the experimenter may provide assistance to participants in completing them. Successful connection to the Wi-Fi and having saved a contact are prerequisites performing subsequent tasks. DIGI-tasks encompass various everyday skills and operations on smart mobile devices. A comprehensive list of all items can be found in Table 1.

Evaluation

An example of an evaluation sheet is displayed in Figure 1. Each item of the DIGI is evaluated based on two major criteria: the selection criterion and production criteria, which correspond to the process of app-selection and of navigating inside the application. The selection criterion pertains to the correct choice of the application suitable for the task (e.g., item *save contact*: input mask is reached (e.g., via contacts, telephone)). The production criteria evaluate the correct solution for each item in two action steps (e.g., for the item *save contact*: 1. Data input, 2. Save). Participants could achieve a maximum of 8 points (tablet: 7) per subtest for selection and 16 points (tablet: 14) for production. The separate evaluation of selection and production criteria is based on the finding that for traditional tools the selection and application can be impaired selectively in stroke patients (Buchmann and Randerath, 2017). Additionally, observable movement-related errors are documented. Typical observable movement errors include imprecise typing, inadequate holding, inadequate pressure, imprecise swiping, and inappropriate zooming. Notably, it is possible to record further movement-related errors. It also needs to be noted that future studies in this field should include kinematics-related evaluation procedures that allow for more precise movement tracking or objective movement error recognition. For each different observed error, one error-point is recorded. For example, if a participant's typing and swiping are both imprecise in one trial, two error-points are noted.

TABLE 1 Overview of the tasks of the DIGI by versions A and B.

DIGI-A	Item	DIGI-B	Item
AB 00.1	Save contact	AB 00.1	Save contact
AB 00.2	Connect to Wi-Fi	AB 00.2	Connect to Wi-Fi
A 01	Charge the device	B 01	Connect the headphones
A 02	Answer a call	B 02	Make a call
A 03	Set an alarm	B 03	Save appointments
A 04	Send smiley	B 04	Send photo
A 05	Mute	B 05	Navigate
A 06	Take a photo	B 06	Zoom in
A 07	Open website	B 07	Set to flight mode
A 08	Zoom out	B 08	Delete photo

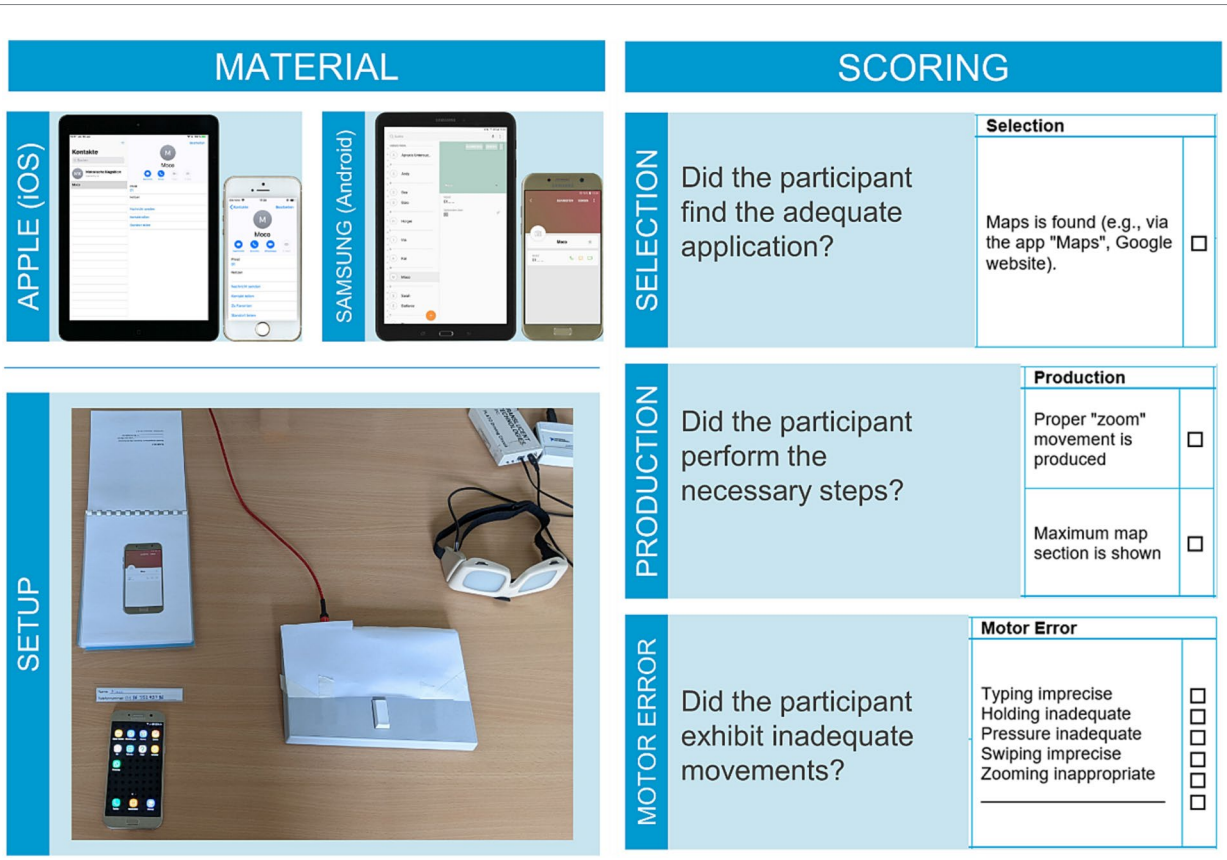


FIGURE 1
Setup and material of the DIGI. The top two left panels show the used devices (smartphones, tablets) and below the DIGI setup is displayed as used in the current study, including response pad and shutter goggles. On the right exemplary excerpts from the evaluation sheets including selection (top), production (middle) and error scores (bottom) are displayed. We would like to point out that the response pad, shutter goggles and head mounted camera are not necessary equipment to conduct the DIGI. At the current developmental stage of the assessment instrument, it served to control the timing of the visual input and the placement of the participant's hand, as well as to refer to video footage of the test sessions whenever necessary (e.g., to test for interrater reliability). For the application in clinical practice, the DIGI requires only the smart mobile devices, booklets, and evaluation sheets.

Material

The material and devices listed below and displayed in Figure 1 were employed for the implementation of the DIGI. The DIGI encompasses form sheets for evaluation, booklets displaying the current task with accompanying photographs showing the target end-state of the device, and paper flashcards with additional information necessary to solve the current task. Further materials include: a multi-socket, device-specific chargers, headphones, an object to be photographed (in this study a toy cat was utilized), and an iOS or Android smartphone and tablet. The smartphones and tablets used in the current study were a Samsung Galaxy A7 smartphone and a Samsung Galaxy Tab A tablet. In our laboratory, the DIGI is also available with an iPhone SE, and an iPad Air. Each device was equipped with a current Android OS or iOS version and received regular updates to ensure optimal functionality.

The primary objective of this study is to introduce the new assessment instrument, DIGI. Consequently, we will focus on the selection, production and motor-error scores. For experimental purposes a Cedrus Response Pad RB-540 was used in the present study. By instructing the participants to press a button of the Response Pad between trials with their hand which operated the digital device, the starting position of the hand was controlled. PLATO Visual Occlusion Spectacles from Translucent Technologies Inc. served to control the timing of visual input. These devices were controlled by a 15.6-inch laptop (ASUS VivoBook) running a Windows 10 Home operating system and the Cedrus Superlab 5 experimental software. To facilitate the evaluation of the participant's performance, a head-mounted camera (GoPro Hero Session) was used to record screen activity during the DIGI.

Procedure

In the course of this study, participants undertook the DIGI assessment using the Android operating system. Each participant completed the test on both, smartphone and tablet.

First, the participant put on the GoPro camera and the goggles. Then two practice trials were conducted followed by the DIGI tasks from versions A and B. The order in which versions and smart mobile devices were presented was balanced evenly among subjects. The response pad was placed adjacent to the hand operating the device. The digital device was placed centrally in front of the participant on the table, showing the home screen. The booklet was placed vertically to the device (see Figure 1). Between the trials, the goggles were shut and the participants placed their hand on the response pad's key.

Each trial started with a verbal instruction of the respective task, consisting of a brief description ("Save contact") and a specification of the task ("Save the number '…' with the name '…' in the contacts"). Participants were given the time they needed, i.e., the task was not time-constrained. Additionally, the booklet with a picture of the successful end-state of the device was presented for reference. This end-state is one of several possible solutions since some items can be solved in various acceptable ways. For example, in devices with Android OS, enabling flight mode may be accomplished via the settings menu – as shown in the booklet. However, it is also possible to enable flight mode via the taskbar, which usually can be dragged down from the upper edge of the home screen. This method results in a visually different, but correct end-state which is credited.

In the present study, the participants were allowed to use their preferred hand or both hands to solve the tasks.

DILA-S

Subtests from the Diagnostic Instrument for Limb Apraxia were administered in this study (DILA-S, for material and manual).[1] The results from the novel (NTT) and familiar (FTT) tools subtests are reported. In both NTT and FTT, participants first select the most appropriate tool from a set of three options and subsequently manipulate an object with the correct tool. The object is either a cylinder (NTT) that shall be lifted from a socket or a well-known everyday object (FTT) that shall be manipulated (e.g., scooping soup from a pot). Participants receive 0–2 selection points per trial, resulting in a total selection score for novel or familiar tool use between 0 and 10 points. Additionally, the participants' ability to correctly manipulate the object is awarded with 0–2 execution-points per trial. This means that the range for the execution-score in NTT and FTT is 0–10 points. For a more comprehensive description of the DILA-S please see Buchmann and Randerath (2017).

Acceptance

The acceptance of the DIGI has been assessed by use of an adapted version of the Akzept! questionnaire by Kersting (2008).

Interrater reliability

Video recordings (received via GoPro) from a subsample of the older group ($n = 7$) were analyzed by a second independent rater who evaluated the participants' performance in the DIGI. Selection, production, and motor-related error scores were summed up across DIGI versions A and B, smartphone, and tablet, and correlated between the experimenter and the independent rater using Kendall's tau.

Data analysis

The normality of the data was assessed on a group-wise basis by using the Kolmogorov–Smirnov test (K-S test). Results of the K-S test indicated that several variables from the DIGI, NTT, and FTT were not normally distributed in either age group ($p < 0.05$). Consequently, the non-parametric Mann–Whitney test was applied for between-group comparisons of DIGI selection-, production-, and error-scores. The Bonferroni-Holm procedure was applied to correct for multiple testing.

1 https://www.moco.uni-konstanz.de/publikationen/assessments/

Correlations between the DIGI selection scores and the selection scores of the NTT and FTT were computed using Kendall's tau. The same procedure was applied for the correlation of DIGI production scores and NTT and FTT execution scores. The correlations were conducted separately for each age group and for each device used.

Results

Analysis of group differences

For an overview of group comparisons concerning the DIGI variables, please see Table 2. Consistent with our hypothesis, we observed that the older age group achieved significantly lower production scores when operating the smartphone or the tablet. Furthermore, the older adults committed significantly more movement-related errors than the young adults on both devices. However, the selection score did not differ significantly between the age groups, for neither smartphone nor tablet (Figure 2).

Correlation of digital and traditional tool use performance (Kendall's tau)

In the younger age group, no significant correlation was identified between the DIGI scores and the performance in the NTT (selection $M = 7.44$; production $M = 19.69$; execution $M = 9.13$) and FTT (selection $M = 9.53$; production $M = 19.60$; execution $M = 9.73$), for both smartphone and tablet ($\tau \leq 0.372$, $p \geq 0.142$). Correlations with the smartphone selection score could not be calculated due to a lack of variance in the younger age group.

Conversely, in the older age group, we observed a significant positive correlation between smartphone production score and NTT execution (execution $M = 8.94$) ($\tau = 0.44$, $p = 0.040$), as well as between tablet production score and NTT execution ($\tau = 0.47$, $p = 0.028$) (Figure 3). Except for these two, there were no further significant correlations between the FTT (selection $M = 10.00$; production $M = 19.75$; execution $M = 9.75$)/NTT (selection $M = 7.75$; production $M = 19.81$) and any of the DIGI variables ($\tau \leq 0.081$, $p \geq 0.728$). Due to a lack of variance, no correlations could be calculated between FTT selection and the DIGI scores in the older age group.

Acceptance

Participants rated the DIGI immediately for acceptance after completing the test items. Mean values and standard deviations for the single items of the acceptability questionnaire are shown in Table 3. There is an overall good acceptance of the DIGI as indicated by both groups. The DIGI has been graded by the older adults with 2.06 (SD = 0.75) and by the younger adults with 1.56 (SD = 0.54) according to the German grading system (1 indicates 'very good' and 6 indicates 'insufficient').

Interrater reliability

We observed significant correlations between the experimenter's and the independent rater's evaluation of the participants' performance in the DIGI on all three scores: Selection score ($\tau = 1.00$, $p < 0.001$), production score ($\tau = 0.781$, $p = 0.015$) and movement-related error scores ($\tau = 0.900$, $p = 0.006$).

Discussion

In the present work, we introduced the DIGI, an assessment tool for evaluating common competencies in handling smartphones and tablets. Through a pilot test involving a small sample of young and older adults, we demonstrated good interrater reliability, feasibility and acceptability of the DIGI assessment. We further showed its potential to detect performance differences in digital tool competencies between younger and older adults.

Our finding suggests that older adults might understand as proficiently as younger adults which application suits best for the assigned task. The older group was able to find and tap the appropriate app-icon on the mobile device and there were no differences between age groups in terms of selection scores.

However, consistent with our hypothesis, the older adults exhibited significantly more problems in producing the correct steps while navigating within the apps. This became evident in group differences for the DIGI production- and movement-related error scores for both smartphone and tablet. Common movement-related errors included, for example, imprecise typing whenever entering text and misperceptions about the meaning of a digital gesture, such as confusing typing and swiping when answering a phone call or confusing the zoom-in and the zoom-out gesture. It seems unlikely

TABLE 2 Comparisons between age groups on DIGI variables by use of the Mann–Whitney test.

Variable	Older (m; SD)	Young (m; SD)	*U*	*z*	*p*	*p* adj.
Smartphone selection score (in %)	98.05;0.3.76	100.00;0	160.00	2.10	0.239	0.478
Smartphone production score (in %)	86.91;7.67	97.46;2.61	241.50	4.35	<0.001	<0.001***
Smartphone error score	2.50;2.66	0.06;0.25	51.00	−3.43	0.003	0.012*
Tablet selection score (in %)	98.21;4.12	99.55;1.79	144.50	1.08	0.539	0.539
Tablet production score (in %)	91.10;8.71	98.66;2.21	201.00	2.93	0.005	0.015*
Tablet error score	3.13;2.85	0.19;0.54	40.00	−3.66	0.001	0.005**
Hand sensibility	2.56;0.79	2.38;0.39	119.50	−0.34	0.752	–

*$p_{adj} \leq 0.05$, **$p_{adj} \leq 0.01$, ***$p_{adj} \leq 0.001$ (adjusted with Bonferroni-Holm procedure).

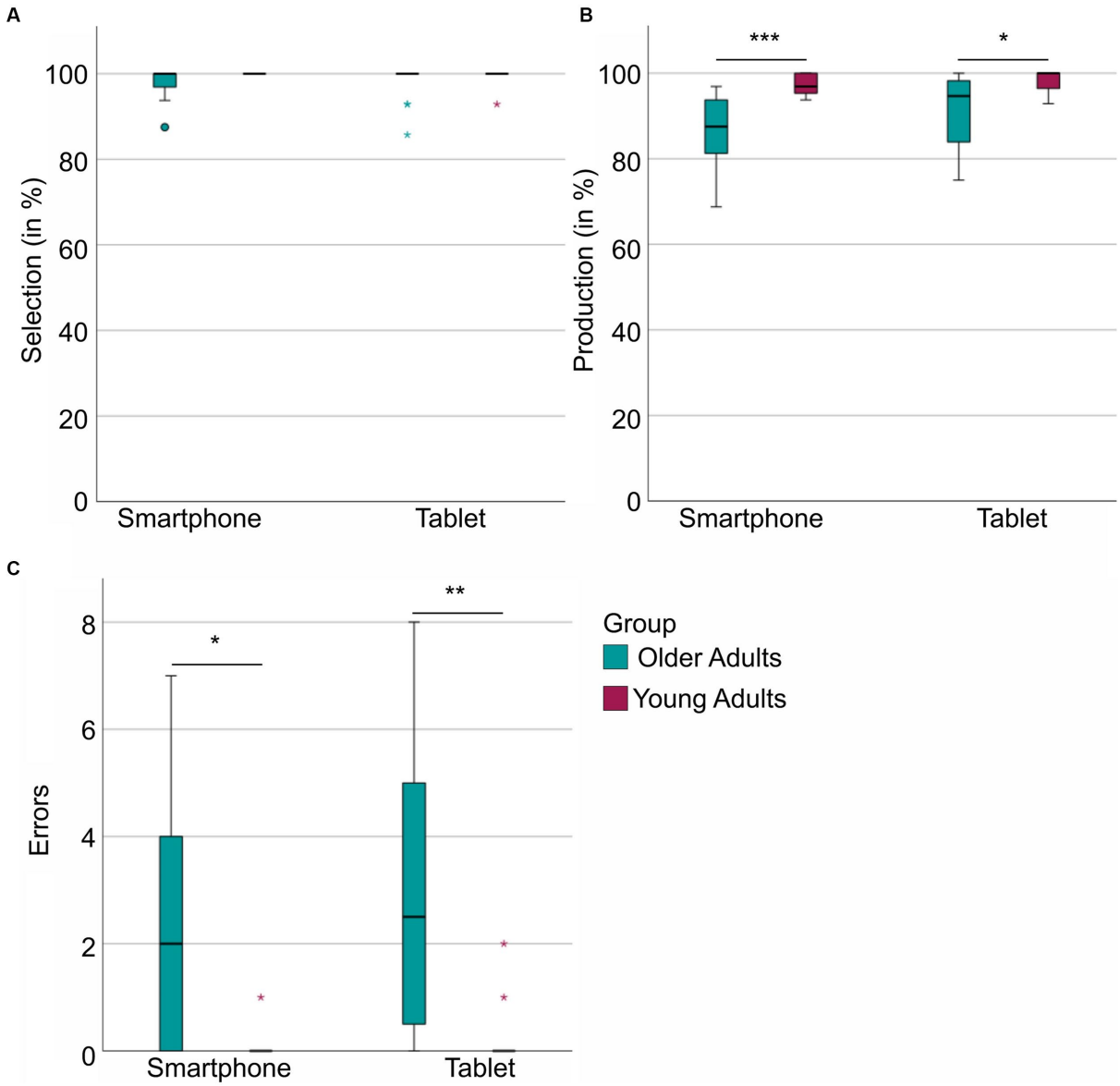


FIGURE 2
DIGI scores per group (older adults vs. young) and per device (smartphone vs. tablet). **(A)** Displays the DIGI selection score in percent per group and device. **(B)** The DIGI production score per group and device. **(C)** The sum of movement-related errors per group and device. *$p_{adj} \leq 0.05$, **$p_{adj} \leq 0.01$, ***$p_{adj} \leq 0.001$ (adjusted with Bonferroni-Holm procedure).

that this deficit could be explained by a decreased hand sensibility in the older age group since we observed that hand sensibility did not differ between groups. The literature, however, demonstrates that older adults show indeed a variety of motor deficits in comparison to younger adults, such as difficulties in coordination, increased variability of movements, slowing of movements, and difficulties with balance and gait, which are attributable to age-related changes in the central nervous system (for an overview see Seidler et al., 2010). These age-related changes in the central nervous system might have contributed to the increase in motor-related errors in the older age group. Early technology-related findings by Smith et al. (1999) support this hypothesis. The authors showed that cursor control tasks with a computer mouse were significantly more difficult for older than younger adults. In their study, this difficulty was associated with age-related declines in motor control, specifically in motor coordination. Comparable mechanisms might have led to the observation of more movement-related errors in the older group of the current study.

Furthermore, in the older age group, the ability to use digital tools correlated with the ability to use traditional (mechanical) but novel tools. Specifically, individuals with lower skills in navigating digital tools tend to display lower skills in applying novel tools to their recipient objects. This could point towards three different interpretations. One hypothesis posits that lower digital tool competencies are indicative of cognitive decline due to healthy aging. Substantiating this hypothesis, the existing literature demonstrates that digital app usage including such characteristics as number of apps used, usage by hour of day, swipes, and keystroke events predicts

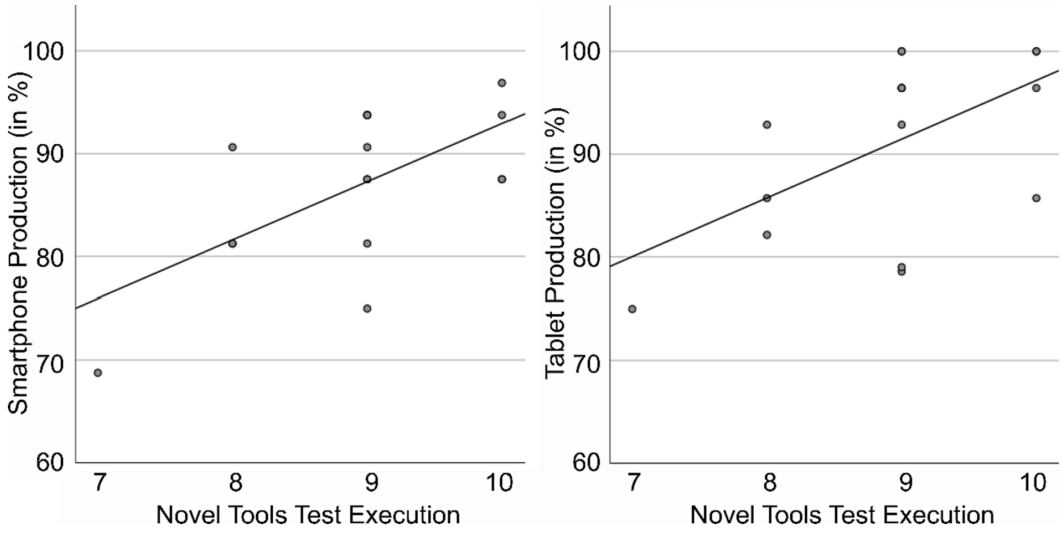


FIGURE 3
Correlation of smartphone and tablet production scores with NTT execution in the older age group.

TABLE 3 Acceptance ratings for the DIGI per group.

	Older adults means (SD)	Young adults means (SD)
Scale: 1 (does not apply) – 6 (applies completely)		
The test tasks were clear and comprehensible.	5.88 (0.34)	5.50 (1.32)
The test can precisely map the differences that exist in relation to the tested characteristic.	5.13 (1.09)	5.13 (1.02)
The test tasks reflect the use of digital devices, which is also required in everyday life.	5.56 (0.89)	4.81 (1.60)
I felt overburdened during the test.	1.56 (1.09)	1.44 (1.26)
It is doubtful that the test will reveal difficulties in the use of digital devices.	2.25 (1.39)	2.31 (1.30)
The test reliably measures what it measures.	5.06 (1.18)	4.50 (1.32)
I did not understand the task.	1.00 (0.00)	1.31 (1.25)
Working through the test tasks is stressful.	1.69 (1.25)	1.38 (1.26)
I always knew what I had to do when working on the test tasks.	5.13 (1.45)	5.00 (1.86)
The ability to perform well in the tested tasks and the ability to use digital devices are two entirely different things.	2.50 (1.51)	2.25 (1.00)
The test allows you to precisely measure the differences in performance between different people in the ability covered by the test.[a]	4.80 (1.08)	4.63 (1.15)
The majority of the test tasks were too difficult for me.	1.19 (0.54)	1.06 (0.25)
The test tasks have too little in common with reality to accurately predict success in the use of digital devices.	1.44 (0.89)	1.69 (0.79)
Working through the test tasks is exhausting.	2.19 (1.80)	1.13 (0.34)
I did not understand the test tasks.	1.06 (0.25)	1.06 (0.25)
The test evaluation can provide an accurate picture of a person's abilities.	4.69 (1.74)	4.81 (0.98)
Scale: 1 (very good) – 6 (insufficient)		
What grade would you give the test you just finished?	2.06 (0.93)	1.56 (0.51)
Compared to other people in my age group (with the same level of education), I think I did … in the test.	2.81 (0.75)	2.19 (0.54)

Items adapted from Akzept! by Kersting (2008) and translated from German. [a]One missing value in the older group for this item (i.e., $n = 15$).

cognitive ability in older adults as measured with neuropsychological assessments (Gordon et al., 2019). The second hypothesis could point towards healthy older people having overall difficulties in novel hand-tool interactions in the sense of mechanical reasoning and thereby showing lower practical digital tool competencies. While previous results (Randerath et al., 2017) suggest that healthy older

versus young subjects do not differ on a group level in performing novel tool use, the current study demonstrates a correlation between novel tool use and digital tool use skills. The third hypothesis directs towards an effect of the cohort with reduced familiarity with digital rules. Older people, who did not grow up surrounded by digital technologies, are sometimes labeled as "digital immigrants" (Prensky, 2001). They may need similar resources to handle digital tools as they need for using traditional (mechanical) novel tools. This may relate to general rule retrieval that is also discussed to be essential for novel tool utilization (Randerath, 2020; Stoll et al., 2022). The Broca area may be a relevant neural correlate that has been associated with different behavioral tasks based on rules, such as rule-guided actions (Bunge, 2004; Donohue et al., 2008), and grammatical rules in language syntax (Tettamanti et al., 2002). An overlap of lesion areas associated with impaired novel tool selection in Broca's area have been discussed to be related to the retrieval and maintenance of object characteristics and physic rules (Stoll et al., 2022). The speculated potential overlap of digital tool competencies and behavioral and neural correlates of rule retrieval and novel tool use needs to be addressed in future studies. The argument that digital immigrants who encountered digital technology much later in life may approach these devices like novel tools is in line with the third hypothesis. Instead, younger people, commonly referred to as digital natives, may use different resources for digital competencies, relying more on common knowledge and overlearned procedures. In accordance with the hypothesis that the brains of digital natives might diverge from those of digital immigrants (Prensky, 2001), we speculate that the younger and the older age groups in our study might have recruited different areas of the brain to solve the DIGI. Participants in the older age group might have employed similar brain regions to solve the DIGI as they do to solve the DILA-S NTT. Our data implies that subsequent studies on the DIGI's psychometric properties need to clarify its underlying constructs cohort-wise as age and the year born may both play a decisive role.

Furthermore, in our study, all participants utilized laboratory-owned smartphones and tablets rather than their personal devices. While this has several practical reasons (standardization, data protection, assessment procedure etc.) there are also some challenges going along with this. For example, a participant might have been familiar with the Android OS in general but running on a Huawei smartphone, and therefore, may not have been versed in its operation on a Samsung device, specifically on a Samsung Galaxy A7 used in the current setup. Dealing with unfamiliar devices can lead to user errors, given variations in the design and operation of different smartphones and tablets (Byrom and Row, 2017; Germine et al., 2019) and perhaps younger participants are more flexible in switching between brands.

It appears notable that the here-described difficulties in handling digital tools in the older sample may further extend to potential non-use of more specific health apps. The question arises of how to secure the inclusion and participation of those suffering from a loss of digital competencies. Digital tool use has gained growing importance not only for the area of health improvement but also in medical diagnostics. For example, current literature discusses approaches that target cognitive digital phenotyping by capturing everyday cognition *in vivo* via digital tool use (Hackett and Giovannetti, 2022). As some studies suggest that app use can predict cognitive performance decline (Gordon et al., 2019), the idea of cognitive digital phenotyping would be, for example, to contribute to early diagnosis of dementia by evaluating a person's app using behavior (Hackett and Giovannetti, 2022). The inevitable growth in these approaches promises increasing gains and advantages but faces many challenges including participation of vulnerable groups.

There are certain methodological limitations and challenges when assessing digital tool competencies such as the handling of smartphones and tablets. It is important to keep the experimental devices in an up-to-date state to ensure their optimal functionality. However, this practice can introduce concerns regarding the comparability of early and later DIGI surveys, since the software, and the UI might change slightly with updates. Similarly, hardware, software, and the way we use it changes rapidly, which may pose a difficulty in the context of the thorough development of a neuropsychological diagnostic instrument (Schmand, 2019). Thus, it is questionable how long the specific tasks included in the DIGI will be relevant for our everyday living. Additionally, it is debatable for how long the specific smartphone−/tablet-brands we included in the DIGI will remain among the most frequently used ones. While we here provide a framework for presenting items and evaluating practical digital competencies of common tasks and features of smartphones and tablets, for future developments, we expect that regular reevaluations and adjustments of the items and devices present necessary steps.

Specific limitations of the current study are the small sample sizes and ceiling effects in certain DIGI and DILA-S variables especially in the young adults group. A major objective for future research is to enlarge our samples for all age groups and incorporate conditions with constraint hand use to obtain control samples for neurologic patients who oftentimes suffer from motor unimanual impairments such as hemiparesis. For our neurologic sample, it will be important to broaden the sample and include more severely impaired patients. The next steps entail collecting psychometric data and evaluating behavioral and neural correlates of diminished digital competencies.

Conclusion

In light of the growing importance of digital devices, we tried to provide one important step towards diagnosing common digital abilities. In the present paper, we introduced an assessment instrument for basic competencies in smartphone and tablet use, the DIGI. We demonstrated its feasibility and acceptability in healthy samples of different ages. Differences between older and younger adults were found particularly for navigation within apps and for producing motor-related errors. Only in older adults worse performance in handling traditional novel tools in the DILA-S went along with reduced digital tool competencies in the DIGI. We speculated that the overlap of digital tool competencies and novel tool use is due to shared correlates of potential rule retrieval.

Follow-up studies should evaluate the DIGI's psychometric properties in larger groups including samples of healthy older participants as well as participants with cognitive impairments such as after suffering from a stroke. To further elucidate the underlying mechanisms of digital tool competencies, future studies should combine behavioral and neuroimaging techniques. When investigating

digital tool competencies it appears particularly important to consider age and year of birth.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Konstanz ethics committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Visualization, Writing – original draft. IB: Data curation, Methodology, Project administration, Writing – review & editing. KH: Data curation, Formal analysis, Visualization, Writing – original draft. JL: Data curation, Methodology, Writing – review & editing. VL: Writing – review & editing, Data curation. FG: Data curation, Writing – review & editing. CH: Data curation, Writing – review & editing, Conceptualization, Methodology. GK: Writing – review & editing. SK: Writing – review & editing. JR: Writing – review & editing, Funding acquisition, Methodology, Resources, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by an intersectional programme of the Zukunftskolleg at the University of Konstanz supported by the Excellence Strategy of the German Federal and State Governments at the University of Konstanz. The Open Access fee was covered by the University of Vienna.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1270437/full#supplementary-material

References

Arab, F., Malik, Y., and Abdulrazak, B. (2013). Evaluation of PhonAge: An Adapted Smartphone Interface for Elderly People. IFIP Conference on Human-Computer Interaction.

Brunzini, A., Caragiuli, M., Atzori, F., Bronzini, M., and Germani, M. (2023). Digital Technology for Elders Better Living: a Usability and User-Experience Assessment. Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments.

Buchmann, I., Dangel, M., Finkel, L., Jung, R., Makhkamova, I., Binder, A., et al. (2020a). Limb apraxia profiles in different clinical samples. *Clin. Neuropsychol.* 34, 217–242. doi: 10.1080/13854046.2019.1585575

Buchmann, I., Finkel, L., Dangel, M., Erz, D., Maren Harscher, K., Kaupp-Merkle, M., et al. (2020b). A combined therapy for limb apraxia and related anosognosia. *Neuropsychol. Rehabil.* 30, 2016–2034. doi: 10.1080/09602011.2019.1628075

Buchmann, I., and Randerath, J. (2017). Selection and application of familiar and novel tools in patients with left and right hemispheric stroke: psychometrics and normative data. *Cortex* 94, 49–62. doi: 10.1016/j.cortex.2017.06.001

Bunge, S. (2004). How we use rules to select actions: a review of evidence from cognitive neuroscience. *Cogn. Affect. Behav. Neurosci.* 4, 564–579. doi: 10.3758/CABN.4.4.564

Busch, M. A., and Kuhnert, R. (2017). 12-Monats-Prävalenz von Schlaganfall Oder Chronischen Beschwerden Infolge Eines Schlaganfalls in Deutschland. *Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung.* doi: 10.17886/RKI-GBE-2017-010

Byrom, B., and Row, B. (2017). The use of digital technologies to collect patient data in outcomes research. *J. Comp. Effec. Res.* 6, 275–277.

Changizi, M., and Kaveh, M. H. (2017). Effectiveness of the mHealth technology in improvement of healthy behaviors in an elderly population – a systematic review. *Mhealth* 3:51. doi: 10.21037/mhealth.2017.08.06

Chiarini, G., Ray, P., Akter, S., Masella, C., and Ganz, A. (2013). mHealth technologies for chronic diseases and elders: a systematic review. *IEEE J. Sel. Areas Commun.* 31, 6–18. doi: 10.1109/JSAC.2013.SUP.0513001

Covello, S., and Lei, J. (2010). A review of digital literacy assessment instruments syracuse university, school of education/IDD & E, IDE-712: Analysis for human performance technology decisions. 1–31.

Donkervoort, M., Dekker, J., Van Den Ende, E., and Stehmann-Saris, J. (2000). Prevalence of apraxia among patients with a first left hemisphere stroke in rehabilitation centres and nursing homes. *Clin. Rehabil.* 14, 130–136. doi: 10.1191/026921500668935800

Donohue, S. E., Wendelken, C., and Bunge, S. A. (2008). Neural correlates of preparation for action selection as a function of specific task demands. *J. Cogn. Neurosci.* 20, 694–706. doi: 10.1162/jocn.2008.20042

Ertl, B., Csanadi, A., and Tarnai, C. (2020). Getting closer to the digital divide: an analysis of impacts on digital competencies based on the German PIAAC sample. *Int. J. Educ. Dev.* 78:102259. doi: 10.1016/j.ijedudev.2020.102259

European Parliament (2006). Recommendation of the European Parliament and the council of 18 December 2006 on key competences for lifelong learning. 394, 10–18. *Off. J. Eur. Union.* Available at: http://data.europa.eu/eli/reco/2006/962/oj

Ferrari, A. (2013). Framework for Developing and Understanding Digital Competence in Europe. *Joint Res. Centre Euro. Comm.* doi: 10.2788/52966

Germine, L., Reinecke, K., and Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *Clin. Neurop.* 33, 271–286.

Goldenberg, G. (2013). Apraxia. Wiley interdisciplinary reviews. *Cogn. Sci.* 4, 453–462. doi: 10.1002/wcs.1241

Gomez-Hernandez, M., Ferre, X., Moral, C., and Villalba-Mora, E. (2023). Design guidelines of Mobile apps for older adults: systematic review and thematic analysis. *JMIR Mhealth Uhealth* 11:e43186. doi: 10.2196/43186

Gordon, M. L., Gatys, L., Guestrin, C., Bigham, J. P., Trister, A., and Patel, K. (2019). App Usage Predicts Cognitive Ability in Older Adults. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Hackett, K., and Giovannetti, T. (2022). Capturing cognitive aging in vivo: application of a neuropsychological framework for emerging digital tools. *JMIR Aging* 5:e38130. doi: 10.2196/38130

Heilman, K. M., and Rothi, L. J. (1993). "Apraxia" in *Clinical neuropsychology*. eds. K. M. Heilman and E. Valenstein (New York, Oxford: Oxford University Press), 141–164.

Hunter, J. M., Schneider, L. H., Mackin, E. J., and Callahan, A. D. (1990). *Rehabilitation of the Hand: Surgery and Therapy, 3rd ed*. St. Louis: Mosby.

Kalbe, E., Kessler, J., Calabrese, P., Smith, R., Passmore, A., Brand, M., et al. (2004). DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. *Int. J. Geriatr. Psychiatry* 19, 136–143. doi: 10.1002/gps.1042

Karnoe, A., Furstrand, D., Christensen, K. B., Norgaard, O., and Kayser, L. (2018). Assessing competencies needed to engage with digital health services: development of the eHealth literacy assessment toolkit. *J. Med. Internet Res.* 20:e8347. doi: 10.2196/jmir.8347

Kersting, M. (2008). Zur akzeptanz von intelligenz-und leistungstests. *Rep. Psychol.* 33, 420–433.

Kortum, P., and Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *Int. J. Hum. Comput. Interact.* 31, 518–529. doi: 10.1080/10447318.2015.1064658

Li, Z., Huang, A., Xu, W., Hu, W., and Xie, L. (2014). Fall Perception for Elderly Care: A Fall Detection Algorithm in Smart Wristlet mHealth System. 2014 IEEE International Conference on Communications (ICC).

Liepmann, H. (1900). Das Krankheitshild der Apraxie ("Motorischen Asymbolie") auf Grund eines Falles von einseitiger Apraxie. *Monatschrift für Psychiatrie und Neurologie* 8, 15–29. doi: 10.1159/000221488

Lu, S.-C., Wenb, T.-N., and Changb, P.-L. (2017). *The Study of Smartphone Usage Competency Assessment and Training for the Elderly* International Medical Informatics Association (IMIA) and IOS Press.

McGaughey, R. E., Zeltmann, S. M., and McMurtrey, M. E. (2013). Motivations and obstacles to smartphone use by the elderly: developing a research framework. *Int. J. Electron. Finance* 7, 177–195. doi: 10.1504/IJEF.2013.058601

Möller, R. (2016). "Das smartphone als Leitmedium" in *Ent-Grenztes Heranwachsen* Wiesbaden: Springer Fachmedien, 185–199. doi: 10.1007/978-3-658-09793-6_10

Oberländer, M., Beinicke, A., and Bipp, T. (2020). Digital competencies: a review of the literature and applications in the workplace. *Comput. Educ.* 146:103752. doi: 10.1016/j.compedu.2019.103752

Prensky, M. (2001). Digital natives, digital immigrants part 2: do they really think differently? *Horizon* 9, 1–6. doi: 10.1108/10748120110424843

Quamar, A. H., Schmeler, M. R., Collins, D. M., and Schein, R. M. (2020). Information communication technology-enabled instrumental activities of daily living: a paradigm shift in functional assessment. *Disabil. Rehabil. Assist. Technol.* 15, 746–753. doi: 10.1080/17483107.2019.1650298

Randerath, J. (2020). *A Simple Illustration of a Left Lateralized Praxis Network: Including a Brief Commentary*, Konstanz, Germany.

Randerath, J. (2023). "Syndromes of limb apraxia: developmental and acquired disorders of skilled movements" in *APA Handbook of Neuropsychology*. eds. G. G. K. Brown, K. Y. Haaland and B. Crosson, vol. *1*. Washington: American Psychological Association. doi: 10.1037/0000307-008

Randerath, J., Buchmann, I., Liepert, J., and Büsching, I. (2017). *Diagnostic Instrument for Limb Apraxia: Short Version (DILA-S), 1st ed*. Konstanz: Universität Konstanz & Lurija Institut.

Sakdulyatham, R., Preeyanont, S., Lipikorn, R., and Watakakosol, R. (2017). User interface on smartphone for elderly users. *Int. J. Autom. Smart Technol.* 7, 147–155. doi: 10.5875/ausmt.v7i4.1339

Schmand, B. (2019). Why are neuropsychologists so reluctant to embrace modern assessment techniques? *Clin. Neuropsychol.* 33, 209–219. doi: 10.1080/13854046.2018.1523468

Seidler, R. D., Bernard, J. A., Burutolu, T. B., Fling, B. W., Gordon, M. T., Gwin, J. T., et al. (2010). Motor control and aging: links to age-related brain structural, functional, and biochemical effects. *Neurosci. Biobehav. Rev.* 34, 721–733. doi: 10.1016/j.neubiorev.2009.10.005

Smith, M. W., Sharit, J., and Czaja, S. J. (1999). Aging, motor control, and the performance of computer mouse tasks. *Hum. Factors* 41, 389–396. doi: 10.1518/001872099779611102

Sonderegger, A., Schmutz, S., and Sauer, J. (2016). The influence of age in usability testing. *Appl. Ergon.* 52, 291–300. doi: 10.1016/j.apergo.2015.06.012

Stoll, S. E., Finkel, L., Buchmann, I., Hassa, T., Spiteri, S., Liepert, J., et al. (2022). 100 years after Liepmann–lesion correlates of diminished selection and application of familiar versus novel tools. *Cortex* 146, 1–23. doi: 10.1016/j.cortex.2021.10.002

Tettamanti, M., Alkadhi, H., Moro, A., Perani, D., Kollias, S., and Weniger, D. (2002). Neural correlates for the acquisition of natural language syntax. *NeuroImage* 17, 700–709. doi: 10.1006/nimg.2002.1201

Wilska, T.-A., and Kuoppamäki, S.-M. (2017). "Necessities to all?: the role of ICTs in the everyday life of the middle-aged and elderly between 1999 and 2014" in *Digital Technologies and Generational Identity* (Routledge), 149–166.

Zhao, X., Wang, L., Ge, C., Zhen, X., Chen, Z., Wang, J., et al. (2020). Smartphone application training program improves smartphone usage competency and quality of life among the elderly in an elder university in China: a randomized controlled trial. *Int. J. Med. Inform.* 133:104010. doi: 10.1016/j.ijmedinf.2019.104010

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 13 February 2024
DOI 10.3389/fpsyg.2024.1369766

OPEN ACCESS

EDITED BY
Alessio Facchin,
University of Milano-Bicocca, Italy

REVIEWED BY
Ciro Rosario Ilardi,
IRCCS SYNLAB SDN, Italy
Nattawan Utoomprurkporn,
Chulalongkorn University, Thailand

*CORRESPONDENCE
Michael Malek-Ahmadi
michael.malekahmadi@bannerhealth.com

RECEIVED 12 January 2024
ACCEPTED 26 January 2024
PUBLISHED 13 February 2024

CITATION
Malek-Ahmadi M and Nikkhahmanesh N (2024) Meta-analysis of Montreal cognitive assessment diagnostic accuracy in amnestic mild cognitive impairment.
Front. Psychol. 15:1369766.
doi: 10.3389/fpsyg.2024.1369766

Meta-analysis of Montreal cognitive assessment diagnostic accuracy in amnestic mild cognitive impairment

Michael Malek-Ahmadi[1,2]* and Nia Nikkhahmanesh[2]

[1]Banner Alzheimer's Institute, Phoenix, AZ, United States, [2]College of Medicine, University of Arizona, Phoenix, AZ, United States

Background: The Montreal Cognitive Assessment (MoCA) is one of the most widely-used cognitive screening instruments and has been translated into several different languages and dialects. Although the original validation study suggested to use a cutoff of ≤26, subsequent studies have shown that lower cutoff values may yield fewer false-positive indications of cognitive impairment. The aim of this study was to summarize the diagnostic accuracy and mean difference of the MoCA when comparing cognitively unimpaired (CU) older adults to those with amnestic mild cognitive impairment (aMCI).

Methods: PubMed and EMBASE databases were searched from inception to 22 February 2022. Meta-analyses for area under the curve (AUC) and standardized mean difference (SMD) values were performed.

Results: Fifty-five observational studies that included 17,343 CU and 8,413 aMCI subjects were selected for inclusion. Thirty-nine studies were used in the AUC analysis while 44 were used in the SMD analysis. The overall AUC value was 0.84 (95% CI: 0.81, 0.87) indicating good diagnostic accuracy and a large effect size was noted for the SMD analysis (Hedge's $g = 1.49$, 95% CI: 1.33, 1.64). Both analyses had high levels of between-study heterogeneity. The median cutoff score for identifying aMCI was <24.

Discussion and conclusion: The MoCA has good diagnostic accuracy for detecting aMCI across several different languages. The findings of this meta-analysis also support the use of 24 as the optimal cutoff when the MoCA is used to screen for suspected cognitive impairment.

KEYWORDS

cognitive screen, mild cognitive impairment, cognitively unimpaired, diagnostic accuracy, cutoff score

Introduction

Amnestic mild cognitive impairment (aMCI) due to Alzheimer's disease (AD) is a syndrome that is associated with future progression to clinical AD. While not all individuals with aMCI progress to AD, they are thought to be at the highest risk of progression and this classification is often referred to as "MCI due to AD" (Albert et al., 2011; Sperling et al., 2011). The diagnostic criteria for aMCI have remained largely the same since their initial publication (Petersen et al., 1999) and require that an individual's episodic memory performance fall at least 1.5 standard

deviations below what would be expected for their age and education level and is accompanied by a self-reported or collateral-reported complaint of cognitive decline. However, the aMCI diagnosis is made only after an extensive neuropsychological examination which prevents the diagnosis from being made in general practice settings where cognitive screening measures are often used to determine if an individual requires a more comprehensive cognitive assessment (Townley et al., 2019). Further refinements to the aMCI diagnostic criteria include the differentiation of those whose impairments are only in the memory domain (single domain) versus those who are impaired in memory and another cognitive domain (multiple domain) (Petersen and Negash, 2008). These classifications also apply for cases where the memory domain is not impaired (non-amnestic MCI), but other domains are (Petersen and Negash, 2008).

For several decades the Mini-Mental State Exam (MMSE) (Folstein et al., 1975) has been the most ubiquitous cognitive screening instrument, however the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) is now among the most widely-used assessments for cognitive screening in general practice settings. Recent evidence indicates that the MoCA is superior to the MMSE in its ability to differentiate aMCI from normal cognition (Pinto et al., 2019a) as many individuals with aMCI often obtain normal scores on the MMSE (26–30) despite collateral reports of significant cognitive decline. The initial validation study of the MoCA recommended the same cutoff score as the MMSE (<26), however subsequent studies have indicated this cutoff may be too stringent and result in false positive indications of possible cognitive impairment (Wong et al., 2015; Carson et al., 2018; Ilardi et al., 2023).

To date, there has not been an extensive review and quantitative analysis of the MoCA's diagnostic accuracy for aMCI. Given that the MoCA has been translated into many different languages and dialects it is important to understand how consistent its diagnostic accuracy is across its various translations. The aims of this meta-analysis are to characterize the MoCA's diagnostic accuracy for aMCI and to characterize its relative effect size for mean differences between cognitively unimpaired (CU) older adults and those with aMCI using a large sample of published observational studies that cover a wide array of the languages that the MoCA has been translated in.

Methods

Inclusion criteria

Prior to conducting the literature searches, the following criteria for study selection and inclusion were established: (1) The data could not come from a treatment or intervention trial, (2) The study should report either raw means and standard deviations for MoCA performance in both the CU and aMCI groups OR the study should report should report area under the curve (AUC) values with standard error (SE) or 95% confidence intervals (CI), (3) The study should use either Petersen criteria (Petersen and Negash, 2008) to classify its aMCI subjects or DSM-V criteria for mild neurocognitive disorder (MND). Although in most circumstances using only one set of diagnostic criteria is preferred, we felt that including studies that used either the Petersen or DSM-V MND criteria would provide greater ecological validity for the study results since the MoCA is used primarily as a screening instrument in general practice settings where formal diagnostic criteria for cognitive impairment are not usually applied. PRISMA guidelines were followed for the analysis and a flow chart depicting study screening and selection is shown in Figure 1.

Literature search terms

Using the PubMed database, four different search terms were used. The first search term, "diagnostic accuracy and Montreal Cognitive Assessment" yielded 686 results from which 38 were screened and 15 were selected for inclusion. A second search using "mild cognitive impairment and Montreal Cognitive Assessment" yielded 1,884 results from which 60 were screened with 26 that were selected for inclusion. The third search using "area under the curve and Montreal Cognitive Assessment" yielded 146 results with 23 that were selected for screening from which five were included. A fourth search using the term "mild neurocognitive disorder and Montreal Cognitive Assessment" did not yield any additional studies beyond those already identified in the previous searches. All four searches were also carried out in the EMBASE database which yielded no additional articles. Nine additional articles were identified through reviews of references sections of the selected papers which brought the final total of included studies to 55 (Figure 1). The search approach taken for this study is consistent with "a multi-faceted approach that uses a series of searches" as described in the *Cochrane Handbook of Systematic Reviews of Interventions* (Lefebvre et al., 2023).

Data quality and extraction

From each of the included studies the following data were extracted: sample sizes for the CU and MCI groups, means and standard deviations of MoCA scores for the CU and MCI groups, AUC values with standard errors (SE). When 95% CIs were reported, SE was derived by taking the difference between the AUC estimate and the upper bound of the 95% CI and dividing by 3.92 (Higgins et al., 2022). The cutpoint associated with the AUC estimate, means and standard deviations for age and education levels (when education was reported in years), and the geographic region in which the study was conducted (Asia, Europe, North America) were also extracted from each study. The quality of each study was assessed using the National Heart, Lung, and Blood Institute (NHLBI) Study Quality Assessment of Case Control Studies[1] which was used to grade each study as Good, Fair, or Poor.

Statistical analysis

The first analytic approach was a meta-analysis of AUC values derived from the receiver operator characteristic (ROC) analyses that differentiated aMCI from CU individuals. The second analytic approach included analyses of the standardized mean difference (SMD) (Hedge's *g*) and the raw mean difference (RMD) for MoCA scores between CU and aMCI. For both analytic approaches, results from random effects analyses

1 https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools

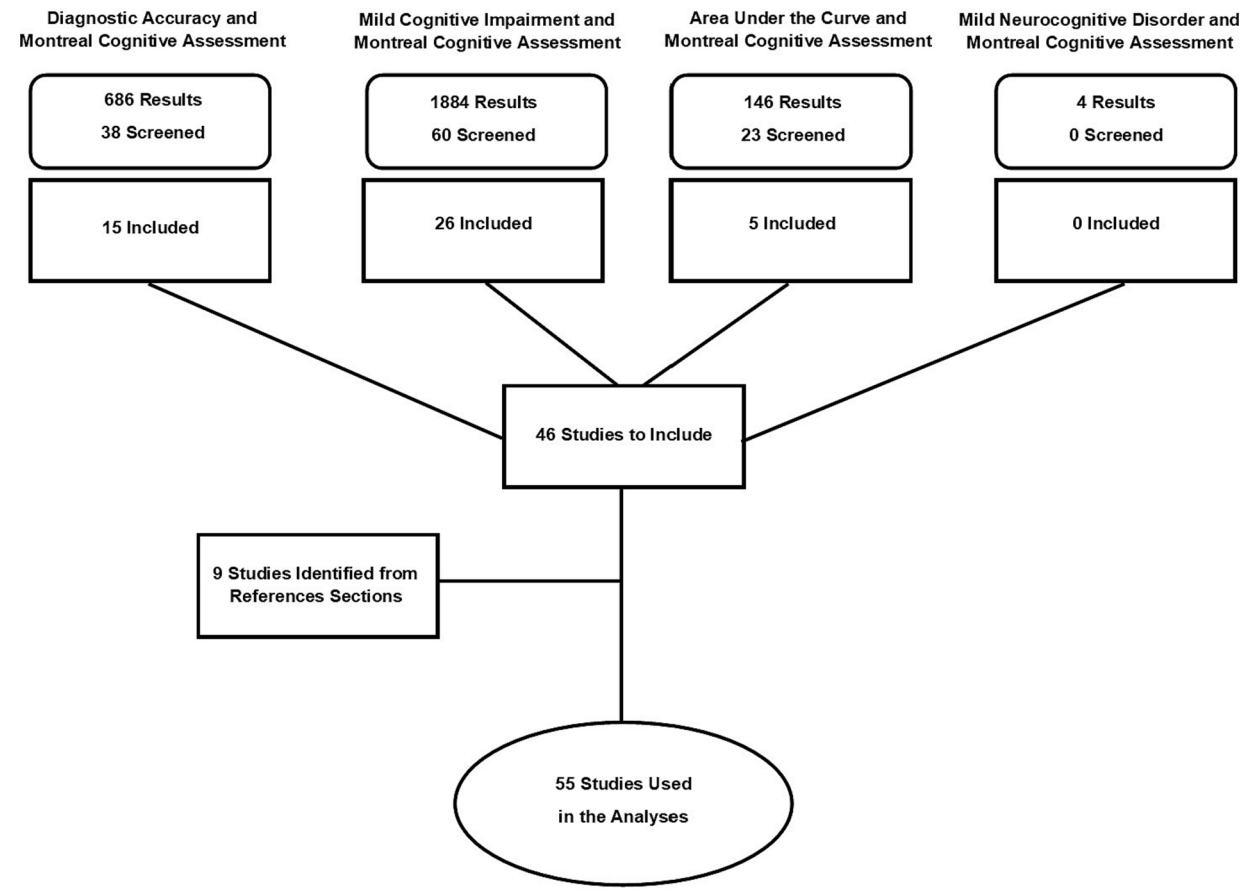


FIGURE 1
PRISMA flowchart for literature searches and number of studies selected for inclusion.

were reported and the I^2 statistic was used to quantify between-study heterogeneity which was classified as low, moderate, or high based on proposed guidelines (Higgins et al., 2003). Additional AUC and SMD analyses were carried out for subgroups based on geographic region (Asia, Europe, North America). The Egger's test was used to determine the presence of publication bias among the included studies. In addition, the median of the reported MoCA cutoff score was used to summarize the reported cutoff values for studies in the AUC analysis. Since the included studies came from a number of different geographic regions we anticipated a wide range of reported MoCA cutoff scores so using the median as a summary measure provides an overall estimate of the MoCA's cutoff that is relatively robust to the variability of reported cutoff values among the studies. All analyses were carried out using MedCalc Statistical Software version 20.109 (MedCalc Software Ltd., Ostend, Belgium,[2] 2022).

TABLE 1 Distribution of MoCA administration language among included studies.

Language of administration	Number of studies
English	15
Mandarin	11
Portuguese	6
Spanish	4
Cantonese	3
Japanese, Turkish	2
Czech, Dutch, Farsi, Georgian, German, Hebrew, Kiswahili, Malay, Mandarin and Malay, Polish, Russian, Thai	1

Results

A total of 55 studies (Fujiwara et al., 2010; Lu et al., 2011; Ahmed et al., 2012; Larner, 2012; Yu et al., 2012; Dong et al., 2013; Freitas et al., 2013; Memória et al., 2013; Roalf et al., 2013; Cummings-Vaughn et al., 2014; Goldstein et al., 2014; Kaya et al., 2014; Malek-Ahmadi et al., 2014; Yeung et al., 2014; Zhou et al., 2014; Chu et al., 2015; Julayanont et al., 2015; Ng et al., 2015; Trzepacz et al., 2015; Mellor et al., 2016; O'Caoimh et al., 2016; Tsai et al., 2016; Cecato et al., 2017; Clarnette et al., 2017; Janelidze et al., 2017; Bartos and Fayette, 2018; Chiu et al., 2018; Lee et al., 2018; Li et al., 2018; Cesar et al., 2019; Delgado et al., 2019; Rossetti et al., 2019; Townley et al., 2019; Wang et al., 2019; Pinto et al., 2019b; Aycicek et al., 2020; Bello-Lepe et al., 2020;

2 https://www.medcalc.org

Dautzenberg et al., 2020; Freud et al., 2020; Senda et al., 2020; Serrano et al., 2020; Sokołowska et al., 2020; Thomann et al., 2020; González et al., 2021; Hemrungrojn et al., 2021; Masika et al., 2021; Rashedi et al., 2021; Rodríguez-Salgado et al., 2021; Yan et al., 2021; Pan et al., 2022; Paterson et al., 2022) were included in this meta-analysis from which 40 were used in the AUC analysis

TABLE 2 Characteristics of studies used in the diagnostic accuracy meta-analysis.

Study	CU sample size	CU age	aMCI sample size	aMCI age	AUC ± SE	Cutoff score
Ahmed et al. (2012)	20	77.4 ± 4.0	15	80.9 ± 7.2	0.89 ± 0.05	23
Cecato et al. (2017)	39	71.6 ± 6.9	44	76.7 ± 7.0	0.93 ± 0.03	24
Clarnette et al. (2017)	41	nr	72	nr	0.94 ± 0.02	23
Cummings-Vaughn et al. (2014)	51	77 ± 7.5	57	78.8 ± 6.7	0.77 ± 0.05	24
Dautzenberg et al. (2020)	459	71.3 ± 7.3	153	73.9 ± 8	0.70 ± 0.02	21
Delgado et al. (2019)	104	72.3 ± 5.4	24	75.3 ± 7.8	0.90 ± 0.03	21
Dong et al. (2013)	128	67.4 ± 4.8	83	74.3 ± 5.5	0.94 ± 0.02	24
Freitas et al. (2013)	90	69.6 ± 7.1	90	70.5 ± 8.0	0.86 ± 0.01	22
Fujiwara et al. (2010)	36	76.4 ± 3.3	30	77.3 ± 6.3	0.95 ± 0.03	25
Goldstein et al. (2014)	16	65.8 ± 7.7	38	71.9 ± 8.9	0.81 ± 0.06	24
Hemrungrojn et al. (2021)	60	67.9 ± 6.4	61	72.1 ± 7.0	0.81 ± 0.04	24
Julayanont et al. (2015)	43	66.6 ± 6.7	42	70.2 ± 6.6	0.90 ± 0.03	25
Kaya et al. (2014)	246	68.0 ± 10.3	114	74.2 ± 8.8	0.85 ± 0.02	nr
Larner (2012)	85	nr	29	nr	0.91 ± 0.02	20
Lee et al. (2018)	35	73.6 ± 6.4	36	76.2 ± 7.4	0.94 ± 0.03	nr
Li et al. (2018)	53	70.2 ± 9.1	56	75.2 ± 7.1	0.82 ± 0.07	24
Liew et al. (2015)	146	64.9 ± 7.0	41	71.8 ± 6.7	0.77 ± 0.05	25
Liu et al. (2021)	50	68.0 ± 8.2	50	76.7 ± 10.8	0.74 ± 0.05	23
Lu et al. (2011)	6,283	72.0	1,687	75.1	0.90 ± 0.005	25
Malek-Ahmadi et al. (2014)	73	82.6 ± 7.7	39	80.5 ± 8.4	0.71 ± 0.05	nr
Masika et al. (2021)	19	69.3 ± 5.8	42	70.4 ± 8.0	0.69 ± 0.07	19
Mellor et al. (2016)	708	72.5 ± 8.4	267	76.5 ± 7.7	0.90 ± 0.01	24
Memória et al. (2013)	28	72.5 ± 5.3	30	74.7 ± 5.7	0.82 ± 0.06	nr
Ng et al. (2015)	88	nr	46	nr	0.50 ± 0.05	nr
O'Caoimh et al. (2016)	101	nr	103	nr	0.84 ± 0.06	24
Pan et al. (2022)	431	66.5 ± 9.3	285	72.1 ± 10.5	0.92 ± 0.01	23
Paterson et al. (2022)	40	74.0 ± 7.0	51	75.0 ± 5.7	0.71 ± 0.05	nr
Peixoto et al. (2018)	30	68.6 ± 6.2	30	67.2 ± 9.3	0.78 ± 0.05	22
Pinto et al. (2019a,b)	110	nr	88	nr	0.95 ± 0.02	nr
Roalf et al. (2013)	140	71.2 ± 9.2	126	72.3 ± 8.1	0.73 ± 0.06	nr
Rodríguez-Salgado et al. (2021)	53	70.4 ± 5.9	46	72.7 ± 7.5	0.73 ± 0.06	nr
Rossetti et al. (2019)	45	62.3 ± 6.8	90	64.8 ± 5.9	0.83 ± 0.04	24
Senda et al. (2020)	50	64.9 ± 12.0	94	73.5 ± 8.3	0.83 ± 0.04	nr
Serrano et al. (2020)	155	71.5 ± 6.2	158	72.6 ± 6.3	0.88 ± 0.02	25
Thomann et al. (2020)	283	73.8 ± 5.2	159	76.0 ± 6.0	0.86 ± 0.01	25
Townley et al. (2019)	313	81.7 ± 5.0	114	84 ± 5.2	0.85 ± 0.02	24
Tsai et al. (2016)	26	nr	59	nr	0.91 ± 0.03	27
Yeung et al. (2014)	49	73.6 ± 7.6	49	76.5 ± 7.5	0.84 ± 0.04	21
Yu et al. (2012)	865	70.4 ± 7.1	115	71.5 ± 7.3	0.71 ± 0.02	21
Zhou et al. (2014)	148	67.7 ± 7.2	24	67.2 ± 6.6	0.72 ± 0.10	26

CU, cognitively unimpaired; MCI, mild cognitive impairment; AUC, area under the curve; SE, standard error; nr, not reported.

and 45 were used in the analysis of mean differences. Thirty-one of the included studies were used in both the AUC and mean difference analyses. An AUC-derived MoCA cutoff score for aMCI was reported by 45 studies. There was a great deal of diversity in the language of administration among the included studies which is shown in Table 1. English was the most prevalent among the studies ($n = 15$) followed by Mandarin ($n = 11$), Portuguese ($n = 6$), and Spanish ($n = 4$). 41% of the included studies were judged to be of good quality while 59% were judged to be of fair quality.

The average age for CU groups was 71.06 ± 7.37 years with an average of 11.44 ± 3.27 years of education. For aMCI groups, the average age was 73.99 ± 7.65 years with an average of 9.89 ± 3.47 years of education. Mean MoCA scores for the CU groups was 24.98 ± 2.88 and 20.11 ± 3.76 for the aMCI groups. Among studies that reported optimal cutoff values ($n = 44$), the median was 24 (range = 17–27). Characteristics of each study included in the AUC meta-analysis are shown in Table 2. The overall AUC value was 0.84, 95% CI (0.81, 0.87), $p < 0.001$ with very high heterogeneity [$I^2 = 90$, 95% CI (87, 92%)] (Figure 2). The Egger's test indicated the presence of publication bias in the analysis ($p = 0.002$). The meta-analysis for differences in means demonstrated a large effect size [Hedge's $g = 1.49$, 95% CI (1.33, 1.64), $p < 0.001$; Table 3] with very high heterogeneity [$I^2 = 93$, 95% CI (91, 94%)] and an Egger's test that indicated the presence of publication bias ($p = 0.003$). The large effect size reported here equates to a 4.73 (95% CI: 4.20, 5.27) point difference on the MoCA between CU and aMCI groups (Figure 3). Characteristics of each study included in the SMD meta-analysis are shown in Table 2. Funnel plots depicting the publication bias in the AUC and SMD analyses are shown in Figure 4.

Analyses of ROC values by geographic region (Table 4) found that North American and Asian studies both yielded AUC values of 0.84 with European studies having a slightly higher AUC of 0.85. For the region-wise SMD analysis (Table 4), Asian studies had the largest effect size [Hedge's $g = 1.67$, 95% CI (1.33, 2.01)], followed by North America [Hedge's $g = 1.23$, 95% CI (1.05, 1.49)] and Europe [Hedge's g = 1.21, 95% CI (0.87, 1.56)].

Discussion

This meta-analysis assessed the diagnostic accuracy and the mean difference of the MoCA when comparing aMCI older adults to those who are CU across several global regions. The overall AUC value of 0.84 indicates that the MoCA has good diagnostic accuracy for aMCI, however a very high degree of between-study heterogeneity was noted for this finding. The analysis of MoCA mean differences yielded a large effect size (Hedge's $g = 1.49$) and a high degree of between-study

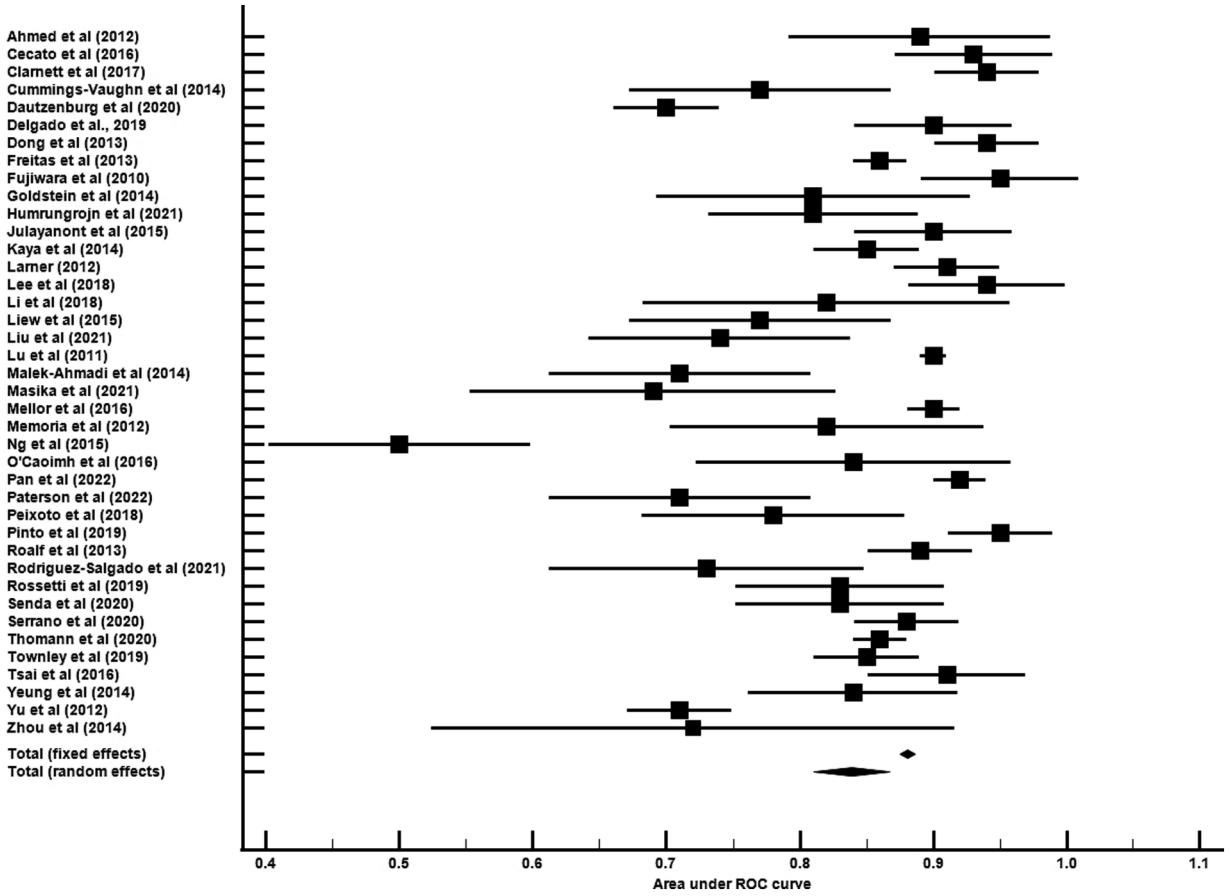


FIGURE 2
Forest plot of MoCA diagnostic accuracy for amnestic mild cognitive impairment.

TABLE 3 Characteristics of studies used in the standardized mean difference meta-analysis.

	Cognitively unimpaired			Mild cognitive impairment		
Study	Sample size	Age	MoCA	Sample size	Age	MoCA
Ahmed et al. (2012)	20	77.4 ± 4.0	27.1 ± 2.8	15	80.9 ± 7.2	21.7 ± 3.3
Aycicek et al. (2020)	91	71.0	22.8 ± 3.2	54	75.0	14.2 ± 5.1
Bartos and Fayette (2018)	226	72.0 ± 8.0	26.0 ± 3.0	48	72.0 ± 7.0	21.0 ± 4.0
Bello-Lepe et al. (2020)	113	71.5 ± 7.6	23.7 ± 3.2	65	76.92 ± 8.71	17.2 ± 4.1
Cesar et al. (2019)	385	nr	19.1 ± 4.9	135	nr	15.1 ± 4.6
Chiu et al. (2018)	99	75.4 ± 6.6	23.1 ± 3.4	128	76.4 ± 6.8	18.3 ± 3.4
Chu et al. (2015)	115	72.2 ± 6.1	24.4 ± 3.2	87	77.2 ± 6.3	18.7 ± 4.6
Cummings-Vaughn et al. (2014)	51	77 ± 7.5	25.8 ± 2.9	57	78.8 ± 6.7	22.8 ± 3.3
Dautzenberg et al. (2020)	459	71.3 ± 7.3	23.5 ± 4.2	153	73.9 ± 8	20.9 ± 3.8
Delgado et al. (2019)	104	72.3 ± 5.4	24.2 ± 3.7	24	75.3 ± 7.8	17.0 ± 3.9
Dong et al. (2013)	128	67.4 ± 4.8	24.3 ± 2.8	83	74.3 ± 5.5	16.4 ± 4.3
Freitas et al. (2013)	90	69.6 ± 7.1	23.6 ± 3.2	90	70.5 ± 8.0	18.3 ± 3.9
Freud et al. (2020)	80	80.1 ± 7.1	24.3 ± 3.7	80	75.0 ± 5.3	20.2 ± 3.1
Goldstein et al. (2014)	16	65.8 ± 7.7	25.1 ± 2.9	38	71.9 ± 8.9	19.8 ± 4.2
González et al. (2021)	3,905	68.0 ± 10.4	26.0 ± 3.0	2,362	70.4 ± 9.0	22.0 ± 4.6
Hemrungrojn et al. (2021)	60	67.9 ± 6.4	28.5 ± 1.8	61	72.1 ± 7.0	26.2 ± 2.2
Janelidze et al. (2017)	46	57.7 ± 10.8	26.3 ± 2.5	20	62.8 ± 11.5	19.2 ± 1.8
Julayanont et al. (2015)	43	66.6 ± 6.7	26.6 ± 1.9	42	70.2 ± 6.6	22.9 ± 2.1
Kaya et al. (2014)	246	68.0 ± 10.3	23.3 ± 3.1	114	74.2 ± 8.8	18.9 ± 3.3
Larner (2012)	85	nr	25.2 ± 3.2	29	nr	18.3 ± 4.5
Lee et al. (2018)	35	73.6 ± 6.4	24.5 ± 2.5	36	76.2 ± 7.4	16.6 ± 5.1
Li et al. (2018)	53	70.2 ± 9.1	25.8 ± 2.3	56	75.2 ± 7.1	20.9 ± 3.3
Liew et al. (2015)	146	64.9 ± 7.0	25.2 ± 2.1	41	71.8 ± 6.7	21.6 ± 4.0
Lifshitz et al. (2012)	80	71.3 ± 4.7	26.7 ± 1.9	74	76.3 ± 5.6	20.3 ± 3.3
Masika et al. (2021)	19	69.3 ± 5.8	20.1 ± 5.4	42	70.4 ± 8.0	15.9 ± 5.9
Mellor et al. (2016)	708	72.5 ± 8.4	27.6 ± 2.7	267	76.5 ± 7.7	21.4 ± 5.5
Memória et al. (2013)	28	72.5 ± 5.3	26.3 ± 2.9	30	74.7 ± 5.7	22.1 ± 3.3
Ng et al. (2015)	88	nr	26.5 ± 3.2	46	nr	26.8 ± 2.7
Pan et al. (2022)	431	66.5 ± 9.3	26.3 ± 3.5	285	72.1 ± 10.5	20.5 ± 5.1
Paterson et al. (2022)	40	74.0 ± 7.0	25.0 ± 2.3	51	75.0 ± 5.7	24.0 ± 2.6
Peixoto et al. (2018)	30	68.6 ± 6.2	26.3 ± 2.5	30	67.2 ± 9.3	21.6 ± 4.9
Rashedi et al. (2021)	59	62.6 ± 6.7	24.5 ± 3.0	40	68.1 ± 8.8	19.3 ± 4.0
Roalf et al. (2013)	140	71.2 ± 9.2	26.8 ± 2.6	126	72.3 ± 8.1	20.9 ± 4.5
Rodríguez-Salgado et al. (2021)	53	70.4 ± 5.9	27.1 ± 2.2	46	72.7 ± 7.5	25.3 ± 2.3
Rossetti et al. (2019)	45	62.3 ± 6.8	25.5 ± 2.1	90	64.8 ± 5.9	21.3 ± 3.9
Senda et al. (2020)	50	64.9 ± 12.0	25.6 ± 2.7	94	73.5 ± 8.3	21.6 ± 3.0
Serrano et al. (2020)	155	71.5 ± 6.2	25.5 ± 2.2	158	72.6 ± 6.3	20.6 ± 3.5
Sokołowska et al. (2020)*	91	74.1	25.9	190	78.2	21.82
Thomann et al. (2020)	283	73.8 ± 5.2	26.5 ± 2.4	159	76.0 ± 6.0	22.0 ± 3.6
Townley et al. (2019)	313	81.7 ± 5.0	24.5 ± 2.5	114	84.0 ± 5.2	20.5 ± 2.9
Trzepacz et al. (2015)	219	77.7 ± 6.2	25.6 ± 2.8	299	74.2 ± 7.9	23.4 ± 3.4
Wang et al. (2019)	136	69.2 ± 11.4	26.5 ± 2.1	120	76.9 ± 7.9	20.2 ± 3.1
Yan et al. (2021)	64	73.5 ± 16.0	26.6 ± 1.0	62	82.0 ± 15.5	20.8 ± 2.7
Yeung et al. (2014)	49	73.6 ± 7.6	22.6 ± 4.0	49	76.5 ± 7.5	16.4 ± 5.0
Yu et al. (2012)	865	70.4 ± 7.11	22.3 ± 5.4	115	71.5 ± 7.3	17.8 ± 6.3
Zhou et al. (2014)	148	67.7 ± 7.2	21.5 ± 0.7	24	67.2 ± 6.6	18.3 ± 1.6

Mean ± standard deviation; nr, not reported. *Standard deviation was not reported.

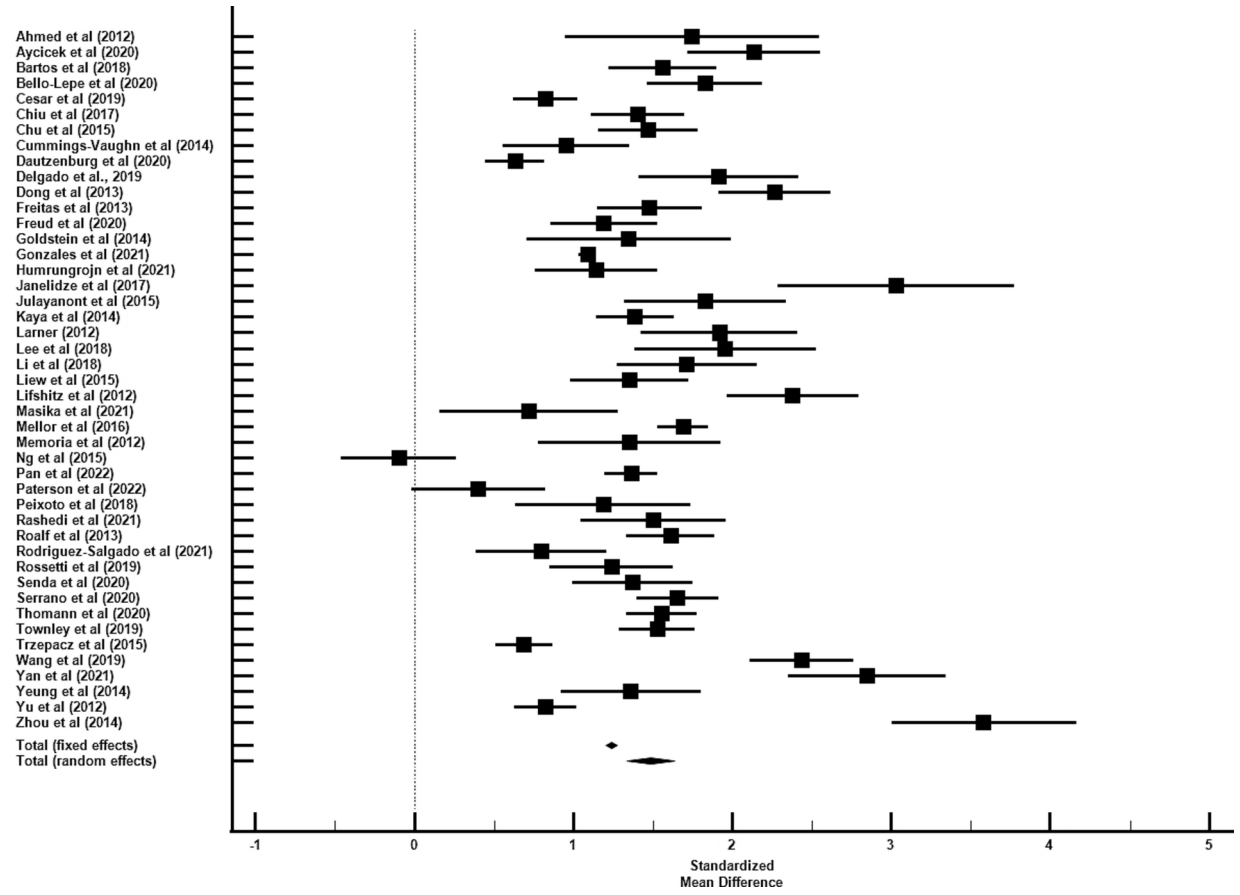


FIGURE 3
Forest plot for standardized mean difference of MoCA performance between CU and aMCI groups.

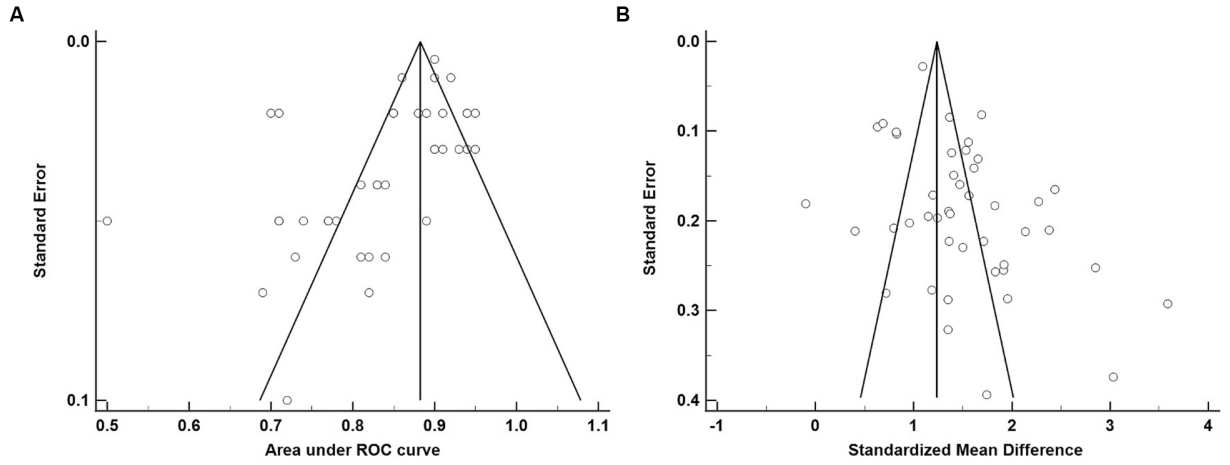


FIGURE 4
Funnel plots for MoCA diagnostic accuracy **(A)** and standardized mean difference **(B)**.

heterogeneity was also noted for this analysis. English and Mandarin studies ($n = 26$) made up approximately half of the studies included in the meta-analysis and among the studies that assessed diagnostic accuracy a score of 24 was the most commonly-used cutoff for differentiating aMCI from CU individuals. However, it was noted that the range of reported cutoff values was 17 to 27 which suggests that optimal MoCA cutpoints may be population- and context-specific in order to avoid misclassification errors. A recent systematic review highlights this point by noting that cross-cultural differences necessitate the use of varying cutoff values as well as corrections for

TABLE 4 Diagnostic accuracy and standardized mean difference analyses stratified by global region.

Global region	AUC (95% CI)	*p*-value	I^2 (95% CI)
North America	0.84 (0.80, 0.88)	<0.001	78% (62, 87%)
Asia	0.84 (0.78, 0.89)	<0.001	93% (89, 95%)
Europe	0.85 (0.79, 0.90)	<0.001	92% (87, 95%)

Global region	Hedge's *g* (95% CI)	*p*-value	I^2 (95% CI)
North America	1.23 (1.05, 1.49)	<0.001	87% (78, 92%)
Asia	1.67 (1.33, 2.01)	<0.001	95% (93, 96%)
Europe	1.21 (0.87, 1.56)	<0.001	90% (82, 95%)

educational levels in different populations (O'Driscoll and Shaikh, 2017).

Others have also noted significant problems with misclassification on the MoCA when a single cutoff is used as higher rates of false positive indications of impairment were noted with increased age and decreased educational levels (Wong et al., 2015). Based on these previous reports the high levels of between-study heterogeneity in this meta-analysis may reflect the cultural, linguistic, and educational diversity among the included studies rather than any particular methodological weakness among them. These findings also emphasize the need to frame the MoCA's utilization in a screening rather than a diagnostic context. Here it is also important to consider the sensitivity and specificity of a cognitive screening measure and how this impacts the utilization of full neuropsychological evaluations. The high false-positive rates of impairment on the MoCA using 26 as the cutoff could lead to many CU individuals being referred for unnecessary neuropsychological evaluations (Ilardi et al., 2023). In contrast, lowering the cutoff score for impairment also has the effect reducing the MoCA's sensitivity in correctly detecting aMCI which further underscores the notion that the MoCA's cutoff score can be adjusted for a given population in order to optimize its diagnostic accuracy. Additionally, adjustments to the cutoff score can be made when physical limitations (e.g., hearing loss) substantially impact MoCA performance (Utoomprurkporn et al., 2020).

A previous meta-analysis of nine studies investigating the MoCA's diagnostic accuracy showed that the optimal MoCA cutoff for detecting aMCI was 23 (Carson et al., 2018) and a recent systematic review found that the AUC value for the MoCA in differentiating aMCI from CU individuals ranged from 0.71 to 0.99 across 34 studies (Pinto et al., 2019b) putting the AUC value of this meta-analysis (0.84) near the midpoint of this range. The three global regions examined in this meta-analysis (North America, Asia, Europe) all had comparable AUC and effect size values despite each region having a high degree of between-study heterogeneity.

There are some limitations to this meta-analysis. Despite the very large number of studies included for both the AUC and SMD analyses, a high degree of between-study heterogeneity was noted for all analyses which decreases the level of confidence one may have in the findings that are reported. A number of different factors may account for the high heterogeneity such as study setting (clinic vs. community-based), varying educational attainment of the populations among the different geographic regions, and cultural norms and values that may impact test performance. While the inconsistencies of the reported AUC and SMD values across studies warrant some degree of skepticism for the final results, there is also significant value in findings that are derived from such a large number of studies across different geographic regions and this aspect of the meta-analysis will likely appeal to clinicians who use the MoCA.

The findings of this meta-analysis provide further support for the use of the MoCA as an accurate cognitive screening tool for use in general practice settings. In line with other studies of the MoCA in aMCI and CU samples, a score of 24 appears to be the optimal cutoff to use for identifying cognitive impairment.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MM-A: Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. NN: Data curation, Formal analysis, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by National Institute on Aging P30AG072980 and P01AG014449.

Conflict of interest

The 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1369766/full#supplementary-material

References

Ahmed, S., de Jager, C., and Wilcock, G. (2012). A comparison of screening tools for the assessment of mild cognitive impairment: preliminary findings. *Neurocase* 18, 336–351. doi: 10.1080/13554794.2011.608365

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Aycicek, G. S., Çalıskan, H., Ozsurekci, C., Unsal, P., Kessler, J., Kalbe, E., et al. (2020). A reliable tool for assessing MCI and dementia: validation study of Dem Tect for Turkish population. *Am. J. Alzheimers Dis. Other Dement.* 35:1533317520949805. doi: 10.1177/1533317520949805

Bartos, A., and Fayette, D. (2018). Validation of the Czech Montreal cognitive assessment for mild cognitive impairment due to Alzheimer disease and Czech norms in 1, 552 elderly persons. *Dement. Geriatr. Cogn. Disord.* 46, 335–345. doi: 10.1159/000494489

Bello-Lepe, S., Alonso-Sánchez, M. F., Ortega, A., Gaete, M., Veliz, M., Lira, J., et al. (2020). Montreal cognitive assessment as screening measure for mild and major neurocognitive disorder in a Chilean population. *Dement Geriatr Cogn Dis Extra.* 10, 105–114. doi: 10.1159/000506280

Carson, N., Leach, L., and Murphy, K. J. (2018). A re-examination of Montreal cognitive assessment (MoCA) cutoff scores. *Int. J. Geriatr. Psychiatry* 33, 379–388. doi: 10.1002/gps.4756

Cecato, J. F., Martinelli, J. E., Izbicki, R., Yassuda, M. S., and Aprahamian, I. (2017). A subtest analysis of the Montreal cognitive assessment (MoCA): which subtests can best discriminate between healthy controls, mild cognitive impairment and Alzheimer's disease? *Int. Psychogeriatr.* 29:701. doi: 10.1017/S104161021600212X

Cesar, K. G., Yassuda, M. S., Porto, F. H. G., Brucki, S. M. D., and Nitrini, R. (2019). MoCA test: normative and diagnostic accuracy data for seniors with heterogeneous educational levels in Brazil. *Arq. Neuropsiquiatr.* 77, 775–781. doi: 10.1590/0004-282x20190130

Chiu, H. F. K., Zhong, B. L., Leung, T., Li, S. W., Chow, P., Tsoh, J., et al. (2018). Development and validation of a new cognitive screening test: the Hong Kong brief cognitive test (HKBC). *Int. J. Geriatr. Psychiatry* 33, 994–999. doi: 10.1002/gps.4883

Chu, L. W., Ng, K. H., Law, A. C., Lee, A. M., and Kwan, F. (2015). Validity of the Cantonese Chinese Montreal cognitive assessment in southern Chinese. *Geriatr Gerontol Int* 15, 96–103. doi: 10.1111/ggi.12237

Clarnette, R., O'Caoimh, R., Antony, D. N., Svendrovski, A., and Molloy, D. W. (2017). (2017). Comparison of the quick mild cognitive impairment (Qmci) screen to the Montreal cognitive assessment (MoCA) in an Australian geriatrics clinic. *Int. J. Geriatr. Psychiatry* 32, 643–649. doi: 10.1002/gps.4505

Cummings-Vaughn, L. A., Chavakula, N. N., Malmstrom, T. K., Tumosa, N., Morley, J. E., and Cruz-Oliver, D. M. (2014). Veterans affairs Saint Louis university mental status examination compared with the Montreal cognitive assessment and the short test of mental status. *J. Am. Geriatr. Soc.* 62, 1341–1346. doi: 10.1111/jgs.12874

Dautzenberg, G., Lijmer, J., and Beekman, A. (2020). Diagnostic accuracy of the Montreal cognitive assessment (MoCA) for cognitive screening in old age psychiatry: determining cutoff scores in clinical practice. Avoiding spectrum bias caused by healthy controls. *Int. J. Geriatr. Psychiatry* 35, 261–269. doi: 10.1002/gps.5227

Delgado, C., Araneda, A., and Behrens, M. I. (2019). Validation of the Spanish-language version of the Montreal cognitive assessment test in adults older than 60 years. Validación del instrumento Montreal cognitive assessment en español en adultos mayores de 60 años. *Neurologia (Engl Ed).* 34, 376–385. doi: 10.1016/j.nrl.2017.01.013

Dong, Y., Yean Lee, W., Hilal, S., Saini, M., Wong, T. Y., Chen, C. L.-H., et al. (2013). Comparison of the Montreal cognitive assessment and the Mini-mental state examination in detecting multi-domain mild cognitive impairment in a Chinese subsample drawn from a population-based study. *Int. Psychogeriatr.* 25, 1831–1838. doi: 10.1017/S1041610213001129

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198. doi: 10.1016/0022-3956(75)90026-6

Freitas, S., Simões, M. R., Alves, L., and Santana, I. (2013). Montreal cognitive assessment: validation study for mild cognitive impairment and Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 27, 37–43. doi: 10.1097/WAD.0b013e3182420bfe

Freud, T., Vostrikov, A., Dwolatzky, T., Punchik, B., and Press, Y. (2020). Validation of the Russian version of the MoCA test as a cognitive screening instrument in cognitively asymptomatic older individuals and those with mild cognitive impairment. *Front Med (Lausanne).* 7:447. doi: 10.3389/fmed.2020.00447

Fujiwara, Y., Suzuki, H., Yasunaga, M., Sugiyama, M., Ijuin, M., Sakuma, N., et al. (2010). Brief screening tool for mild cognitive impairment in older Japanese: validation of the Japanese version of the Montreal cognitive assessment. *Geriatr Gerontol Int* 10, 225–232. doi: 10.1111/j.1447-0594.2010.00585.x

Goldstein, F. C., Ashley, A. V., Miller, E., Alexeeva, O., Zanders, L., and King, V. (2014). Validity of the Montreal cognitive assessment as a screen for mild cognitive impairment and dementia in African Americans. *J. Geriatr. Psychiatry Neurol.* 27, 199–203. doi: 10.1177/0891988714524630

González, D. A., Gonzales, M. M., Jennette, K. J., Soble, J. R., and Fongang, B. (2021). Cognitive screening with functional assessment improves diagnostic accuracy and attenuates bias. *Alzheimers Dement (Amst).* 13:e12250. doi: 10.1002/dad2.12250

Hemrungrojn, S., Tangwongchai, S., Charoenboon, T., Panasawat, M., Supasitthumrong, T., Chaipreserstud, P., et al. (2021). Use of the Montreal cognitive assessment Thai version to discriminate amnestic mild cognitive impairment from Alzheimer's disease and healthy controls: machine learning results. *Dement. Geriatr. Cogn. Disord.* 50, 183–194. doi: 10.1159/000517822

Higgins, JPT, Thomas, J, Chandler, J, Cumpston, M, Li, T, Page, MJ, et al. Cochrane handbook for systematic reviews of interventions version 6.3 (updated February 2022). (2022). Available at: www.training.cochrane.org/handbook.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ* 327, 557–560. doi: 10.1136/bmj.327.7414.557

Ilardi, C. R., Menichelli, A., Michelutti, M., Cattaruzza, T., and Manganotti, P. (2023). Optimal MoCA cutoffs for detecting biologically-defined patients with MCI and early dementia. *Neurol. Sci.* 44, 159–170. doi: 10.1007/s10072-022-06422-z

Janelidze, M., Mikeladze, N., Bochorishvili, N., Dzagnidze, A., Kapianidze, M., Mikava, N., et al. (2017). Validity of the Georgian Montreal cognitive assessment for the screening of mild cognitive impairment and dementia. *Am. J. Alzheimers Dis. Other Dement.* 32, 36–40. doi: 10.1177/1533317516679304

Julayanont, P., Tangwongchai, S., Hemrungrojn, S., Tunvirachaisakul, C., Phanthumchinda, K., Hongsawat, J., et al. (2015). The Montreal cognitive assessment-basic: a screening tool for mild cognitive impairment in illiterate and low-educated elderly adults. *J. Am. Geriatr. Soc.* 63, 2550–2554. doi: 10.1111/jgs.13820

Kaya, Y., Aki, O. E., Can, U. A., Derle, E., Kibaroğlu, S., and Barak, A. (2014). Validation of Montreal cognitive assessment and discriminant power of Montreal cognitive assessment subtests in patients with mild cognitive impairment and Alzheimer dementia in Turkish population. *J. Geriatr. Psychiatry Neurol.* 27, 103–109. doi: 10.1177/0891988714522701

Larner, A. J. (2012). Screening utility of the Montreal cognitive assessment (MoCA): in place of--or as well as--the MMSE? *Int. Psychogeriatr.* 24, 391–396. doi: 10.1017/S1041610211001839

Lee, M. T., Chang, W. Y., and Jang, Y. (2018). Psychometric and diagnostic properties of the Taiwan version of the quick mild cognitive impairment screen. *PLoS One* 13:e0207851. doi: 10.1371/journal.pone.0207851

Lefebvre, C., Glanville, J., Briscoe, S., Featherstone, R., Littlewood, A., Metzendorf, M.-I., et al. (2023). "Chapter 4: searching for and selecting studies" in *Cochrane handbook for systematic reviews of interventions version 6.4*. eds. H. JPT, J. Thomas, J. Chandler, M. Cumpston, T. Li and M. J. Page et al. (Cochrane).

Li, X., Jia, S., Zhou, Z., Zhang, X., Zheng, W., Rong, P., et al. (2018). The role of the Montreal cognitive assessment (MoCA) and its memory tasks for detecting mild cognitive impairment. *Neurol. Sci.* 39, 1029–1034. doi: 10.1007/s10072-018-3319-0

Liew, T. M., Feng, L., Gao, Q., Ng, T. P., and Yap, P. (2015). Diagnostic utility of Montreal cognitive assessment in the fifth edition of diagnostic and statistical manual of mental disorders: major and mild neurocognitive disorders. *J. Am. Med. Dir. Assoc.* 16, 144–148. doi: 10.1016/j.jamda.2014.07.021

Lifshitz, M., Dwolatzky, T., and Press, Y. (2012). Validation of the Hebrew version of the MoCA test as a screening instrument for the early detection of mild cognitive impairment in elderly individuals. *J. Geriatr. Psychiatry Neurol.* 25, 155–161. doi: 10.1177/0891988712457047

Liu, X., Chen, X., Zhou, X., Shang, Y., Xu, F., Zhang, J., et al. (2021). Validity of the mem Trax memory test compared to the Montreal cognitive assessment in the detection of mild cognitive impairment and dementia due to Alzheimer's disease in a Chinese cohort. *J. Alzheimers Dis.* 80, 1257–1267. doi: 10.3233/JAD-200936

Lu, J., Li, D., Li, F., Zhou, A., Wang, F., Zuo, X., et al. (2011). Montreal cognitive assessment in detecting cognitive impairment in Chinese elderly individuals: a population-based study. *J. Geriatr. Psychiatry Neurol.* 24, 184–190. doi: 10.1177/0891988711422528

Malek-Ahmadi, M., Davis, K., Belden, C. M., and Sabbagh, M. N. (2014). Comparative analysis of the Alzheimer questionnaire (AQ) with the CDR sum of boxes, MoCA, and MMSE. *Alzheimer Dis. Assoc. Disord.* 28, 296–298. doi: 10.1097/WAD.0b013e3182769731

Masika, G. M., Yu, D. S. F., and Li, P. W. C. (2021). Accuracy of the Montreal cognitive assessment in detecting mild cognitive impairment and dementia in the rural African population. *Arch. Clin. Neuropsychol.* 36, 371–380. doi: 10.1093/arclin/acz086

Mellor, D., Lewis, M., McCabe, M., Byrne, L., Wang, T., Wang, J., et al. (2016). Determining appropriate screening tools and cut-points for cognitive impairment in an elderly Chinese sample. *Psychol. Assess.* 28, 1345–1353. doi: 10.1037/pas0000271

Memória, C. M., Yassuda, M. S., Nakano, E. Y., and Forlenza, O. V. (2013). Brief screening for mild cognitive impairment: validation of the Brazilian version of the

Montreal cognitive assessment. *Int. J. Geriatr. Psychiatry* 28, 34–40. doi: 10.1002/gps.3787

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Ng, T. P., Feng, L., Lim, W. S., Chong, M. S., Lee, T. S., Yap, K. B., et al. (2015). Montreal cognitive assessment for screening mild cognitive impairment: variations in test performance and scores by education in Singapore. *Dement. Geriatr. Cogn. Disord.* 39, 176–185. doi: 10.1159/000368827

O'Caoimh, R., Timmons, S., and Molloy, D. W. (2016). Screening for mild cognitive impairment: comparison of "MCI specific" screening instruments. *J. Alzheimers Dis.* 51, 619–629. doi: 10.3233/JAD-150881

O'Driscoll, C., and Shaikh, M. (2017). Cross-cultural applicability of the Montreal cognitive assessment (MoCA): a systematic review. *J. Alzheimers Dis.* 58, 789–801. doi: 10.3233/JAD-161042

Pan, F. F., Cui, L., Li, Q. J., and Guo, Q. H. (2022). Validation of a modified Chinese version of Mini-Addenbrooke's cognitive examination for detecting mild cognitive impairment. *Brain Behav.* 12:e2418. doi: 10.1002/brb3.2418

Paterson, T. S. E., Sivajohan, B., Gardner, S., Binns, M. A., Stokes, K. A., Freedman, M., et al. (2022). Accuracy of a self-administered online cognitive assessment in detecting amnestic mild cognitive impairment. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 77, 341–350. doi: 10.1093/geronb/gbab097

Peixoto, B., Machado, M., Rocha, P., Macedo, C., Machado, A., Baeta, E., et al. (2018). Validation of the Portuguese version of Addenbrooke's cognitive examination III in mild cognitive impairment and dementia. *Adv. Clin. Exp. Med.* 27, 781–786. doi: 10.17219/acem/68975

Petersen, R. C., and Negash, S. (2008). Mild cognitive impairment. *CNS Spectr.* 13, 46–53. doi: 10.1017/s1092852900016151

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Arch. Neurol.* 56, 303–308. doi: 10.1001/archneur.56.3.303

Pinto, T. C. C., Machado, L., Bulgacov, T. M., Rodrigues-Júnior, A. L., Costa, M. L. G., Ximenes, R. C. C., et al. (2019a). Is the Montreal cognitive assessment (MoCA) screening superior to the Mini-mental state examination (MMSE) in the detection of mild cognitive impairment (MCI) and Alzheimer's disease (AD) in the elderly? *Int. Psychogeriatr.* 31, 491–504. doi: 10.1017/S1041610218001370

Pinto, T. C. C., Machado, L., Costa, M. L. G., Santos, M. S. P., Bulgacov, T. M., Rolim, A. P. P., et al. (2019b). Accuracy and psychometric properties of the Brazilian version of the Montreal cognitive assessment as a brief screening tool for mild cognitive impairment and Alzheimer's disease in the initial stages in the elderly. *Dement. Geriatr. Cogn. Disord.* 47, 366–374. doi: 10.1159/000501308

Rashedi, V., Foroughan, M., and Chehrehnegar, N. (2021). Psychometric properties of the Persian Montreal cognitive assessment in mild cognitive impairment and Alzheimer disease. *Dement. Geriatr. Cogn. Dis. Extra.* 11, 51–57. doi: 10.1159/000514673

Roalf, D. R., Moberg, P. J., Xie, S. X., Wolk, D. A., Moelter, S. T., and Arnold, S. E. (2013). Comparative accuracies of two common screening instruments for classification of Alzheimer's disease, mild cognitive impairment, and healthy aging. *Alzheimers Dement.* 9, 529–537. doi: 10.1016/j.jalz.2012.10.001

Rodríguez-Salgado, A. M., Llibre-Guerra, J. J., Tsoy, E., Penalver-Guia, A. I., Bringas, G., Erlhoff, S. J., et al. (2021). A brief digital cognitive assessment for detection of cognitive impairment in Cuban older adults. *J. Alzheimers Dis.* 79, 85–94. doi: 10.3233/JAD-200985

Rossetti, H. C., Smith, E. E., Hynan, L. S., Lacritz, L. H., Cullum, C. M., Van Wright, A., et al. (2019). Detection of mild cognitive impairment among community-dwelling African Americans using the Montreal cognitive assessment. *Arch. Clin. Neuropsychol.* 34, 809–813. doi: 10.1093/arclin/acy091

Senda, M., Terada, S., Takenoshita, S., Hayashi, S., Yabe, M., Imai, N., et al. (2020). Diagnostic utility of the Addenbrooke's cognitive examination-III (ACE-III), Mini-ACE, Mini-mental state examination, Montreal cognitive assessment, and Hasegawa dementia scale-revised for detecting mild cognitive impairment and dementia. *Psychogeriatrics* 20, 156–162. doi: 10.1111/psyg.12480

Serrano, C. M., Sorbara, M., Minond, A., Finlay, J. B., Arizaga, R. L., Iturry, M., et al. (2020). Validation of the argentine version of the Montreal cognitive assessment test (MOCA): a screening tool for mild cognitive impairment and mild dementia in elderly. *Dement Neuropsychol.* 14, 145–152. doi: 10.1590/1980-57642020dn14-020007

Sokołowska, N., Sokołowski, R., Oleksy, E., Kasperska, P., Klimkiewicz-Wszelaki, K., Polak-Szabela, A., et al. (2020). Usefulness of the polish versions of the Montreal cognitive assessment 7.2 and the Mini-mental state examination as screening instruments for the detection of mild neurocognitive disorder. *Neurol. Neurochir. Pol.* 54, 440–448. doi: 10.5603/PJNNS.a2020.0064

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on aging Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement.* 7, 280–292. doi: 10.1016/j.jalz.2011.03.003

Thomann, A. E., Berres, M., Goettel, N., Steiner, L. A., and Monsch, A. U. (2020). Enhanced diagnostic accuracy for neurocognitive disorders: a revised cut-off approach for the Montreal cognitive assessment. *Alzheimers Res. Ther.* 12:39. doi: 10.1186/s13195-020-00603-8

Townley, R. A., Syrjanen, J. A., Botha, H., Kremers, W. K., Aakre, J. A., Fields, J. A., et al. (2019). Comparison of the short test of mental status and the Montreal cognitive assessment across the cognitive Spectrum. *Mayo Clin. Proc.* 94, 1516–1523. doi: 10.1016/j.mayocp.2019.01.043

Trzepacz, P. T., Hochstetler, H., Wang, S., Walker, B., and Saykin, A. J.Alzheimer's Disease Neuroimaging Initiative (2015). Relationship between the Montreal cognitive assessment and Mini-mental state examination for assessment of mild cognitive impairment in older adults. *BMC Geriatr.* 15:107. doi: 10.1186/s12877-015-0103-3

Tsai, J. C., Chen, C. W., Chu, H., Yang, H.-L., Chung, M.-H., Liao, Y.-M., et al. (2016). Comparing the sensitivity, specificity, and predictive values of the Montreal cognitive assessment and Mini-mental state examination when screening people for mild cognitive impairment and dementia in Chinese population. *Arch. Psychiatr. Nurs.* 30, 486–491. doi: 10.1016/j.apnu.2016.01.015

Utoomprurkporn, N., Woodall, K., Stott, J., Costafreda, S. G., and Bamiou, D. E. (2020). Hearing-impaired population performance and the effect of hearing interventions on Montreal cognitive assessment (MoCA): systematic review and meta-analysis. *Int. J. Geriatr. Psychiatry* 35, 962–971. doi: 10.1002/gps.5354

Wang, B. R., Zheng, H. F., Xu, C., Sun, Y., Zhang, Y. D., and Shi, J. Q. (2019). Comparative diagnostic accuracy of ACE-III and MoCA for detecting mild cognitive impairment. *Neuropsychiatr. Dis. Treat.* 15, 2647–2653. doi: 10.2147/NDT.S212328

Wong, A., Law, L. S., Liu, W., Wang, Z., Lo, E. S. K., Lau, A., et al. (2015). Montreal cognitive assessment: one cutoff never fits all. *Stroke* 46, 3547–3550. doi: 10.1161/STROKEAHA.115.011226

Yan, M., Yin, H., Meng, Q., Wang, S., Ding, Y., Li, G., et al. (2021). A virtual supermarket program for the screening of mild cognitive impairment in older adults: diagnostic accuracy study. *JMIR Serious Games.* 9:e30919. doi: 10.2196/30919

Yeung, P. Y., Wong, L. L., Chan, C. C., Leung, J. L., and Yung, C. Y. (2014). A validation study of the Hong Kong version of Montreal cognitive assessment (HK-MoCA) in Chinese older adults in Hong Kong. *Hong Kong Med. J.* 20, 504–510. doi: 10.12809/hkmj144219

Yu, J., Li, J., and Huang, X. (2012). The Beijing version of the Montreal cognitive assessment as a brief screening tool for mild cognitive impairment: a community-based study. *BMC Psychiatry* 12:156. doi: 10.1186/1471-244X-12-156

Zhou, S., Zhu, J., Zhang, N., Wang, B., Li, T., Lv, X., et al. (2014). The influence of education on Chinese version of Montreal cognitive assessment in detecting amnesic mild cognitive impairment among older people in a Beijing rural community. *Sci. World J.* 2014:689456. doi: 10.1155/2014/689456

frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 07 May 2024
DOI 10.3389/fpsyg.2024.1394483

OPEN ACCESS

EDITED BY
Alessio Facchin,
Magna Graecia University, Italy

REVIEWED BY
Ottavia Maddaluno,
Santa Lucia Foundation (IRCCS), Italy
Sara B. Festini,
University of Tampa, United States

*CORRESPONDENCE
Odelia Elkana
✉ Odelia.elkana@gmail.com;
✉ elkana@mta.ac.il

RECEIVED 01 March 2024
ACCEPTED 12 April 2024
PUBLISHED 07 May 2024

CITATION
Elkana O (2024) Navigating the "frontal lobe paradox": integrating Real-Life Tasks (RLTs) approach into neuropsychological evaluations. *Front. Psychol.* 15:1394483. doi: 10.3389/fpsyg.2024.1394483

Navigating the "frontal lobe paradox": integrating Real-Life Tasks (RLTs) approach into neuropsychological evaluations

Odelia Elkana[1,2]*

[1]Behavioral Sciences, Academic College of Tel Aviv-Yaffo, Tel Aviv, Israel, [2]The National Institute of Neuropsychological Rehabilitation, Tel Aviv, Israel

KEYWORDS

frontal lobe paradox, frontal lobe dysfunction, executive function, dysexecutive syndrome, prefrontal lobe, neuropsychological assessment, Real-Life Task (RLT), RLT

Introduction

Individuals with frontal lobe damage often exhibit proficiency in interviews and standardized assessment tests, while experiencing significant impairments in their daily functioning—an intriguing phenomenon known as the "frontal lobe paradox" (Stuss and Benson, 1984; Burgess et al., 2006, 2009; Worthington, 2019; Fisher-Hicks et al., 2021; Newstead et al., 2022).

Within the subset of patients with prefrontal cortex (PFC) damage, there is a notable competence observed during clinical interviews and traditional assessments. However, these individuals frequently demonstrate substantial limitations in adaptive functioning, contributing to the complexity of the "frontal lobe paradox" (Stuss and Benson, 1984; Walsh, 1985) or the "knowing-doing dissociation" (Teuber, 1964; Luria, 1980). This not only challenges the clinician's understanding but also places the neuropsychologist in a predicament, as they must grapple with explaining this discrepancy or, in extreme cases, ignore test results that do not align with their diagnosis conclusion. Moreover, failure to address this discrepancy during standardized neuropsychological assessments can have profound consequences for patients, potentially impeding their access to necessary care and supervision, and even exposing them to risks (Fisher-Hicks et al., 2021).

Wood and Bigler (2017) emphasize the significance of conducting comprehensive interviews with individuals who have direct insight into the person's real-world behavior over time to avoid forming misguided opinions solely based on test performance. Burgess et al. (2009) further note that these patients may articulate plans and recall their actions but ultimately struggle to execute intended tasks.

Although many neuropsychologists are familiar with the "frontal lobe paradox", it is common to face challenges in identifying such impairment solely based on standardized test results in typical clinical settings. George and Gilbert (2018) discuss these challenges in relation to the "frontal lobe paradox", addressing the limitations of existing assessment tools and providing insights into the factors contributing to successful performance on standard tests.

To address the challenges posed by the "frontal lobe paradox" (Burgess et al., 2006; Wood and Bigler, 2017; Worthington, 2019; Fisher-Hicks et al., 2021) and ensure comprehensive and valid neuropsychological evaluations, it is imperative to incorporate ecological validity assessment into the assessment process (Goldstein and Scheerer, 1941; Burgess et al., 2006; Fisher-Hicks et al., 2021). Such assessment involves the evaluation of individuals' abilities in real-world contexts, providing valuable insights into their functional abilities and adaptive behaviors in everyday life settings. By supplementing traditional standardized tests with ecological validity measures, clinicians can gain

a more holistic understanding of patients' cognitive functioning and identify discrepancies between performance in controlled testing environments and real-life situations. This integrative approach allows for a more nuanced assessment of executive functioning and self-initiation, particularly in individuals with frontal lobe damage who may demonstrate a disconnect between their performance on standardized tests and their functional abilities in daily life.

The following sections will explore various domains of executive functions along with corresponding RLT examples. This is intended to stimulate further consideration rather than presenting a definitive protocol for integrating RLT into neuropsychological assessment.

Proposed evaluation approach for frontal lobe dysfunction—integrating "Real-Life Tasks" (RLT)

Task initiation and execution of goal-directed behaviors

Executive functioning deficits, particularly in task initiation, are commonly observed in individuals with damage to the prefrontal cortex (PFC) (Stuss and Benson, 1984). Despite intact cognitive abilities measured by traditional neuropsychological tests (Lezak et al., 2012; Goldstein et al., 2013), these individuals often struggle with initiating and executing goal-directed behaviors. This challenge becomes more pronounced in unstructured tasks, where the individual must rely on internal cues and self-initiation to begin and complete activities. Therefore, assessing task initiation abilities within the context of daily life activities is crucial, as it provides valuable insights into individuals' functional capacities and adaptive behaviors.

RLT Example: Present the participant with unstructured tasks (e.g., making a coffee, organizing a desk). Instruct them to start each task without specific guidance. Assess their ability to initiate tasks without external cues and prompts.

To further illustrate, let's delve into the coffee-making example. The participant is asked to make a cup of coffee. Initially, the clinician observes whether the participant asks for directions or clarification, such as the location of the kitchen or where to find the necessary utensils. Then, as the participant progresses through the task, the clinician observes how they navigate each step of the process, from boiling water to selecting and adding sugar or a sugar substitute, choosing and adding milk or a milk substitute, and finally, locating a spoon to mix the coffee. The participant's ability to initiate each step of the task without external guidance is evaluated, along with their overall proficiency in completing the task independently.

Behavioral organization in non-routine situations

Individuals with frontal lobe damage often struggle with planning, organizing, and adapting to novel or complex tasks, indicative of executive functioning deficits (Gioia et al., 1996; Burgess et al., 2006; Lezak et al., 2012).

RLT Example: Create a scenario requiring the participant to plan a social gathering such as a dinner party or a family barbecue given specific event details such as guest count, dietary restrictions, and budget constraints. They must then devise a detailed plan, covering menu selection, ingredient shopping, meal preparation, and venue setup.

During the task, the clinician observes how the participant organizes and prioritizes tasks, allocates resources (time, money), and handles potential challenges. The participants' written plan provides insight into their organizational strategies.

To further evaluate behavioral organization skills, the clinician can assess:

- **Menu planning**: does the participant create a balanced menu considering guest preferences and dietary needs, along with cost-effectiveness and ease of preparation?
- **Budget management**: how effectively does the participant allocate the budget to different event aspects, staying within constraints?
- **Time management**: does the participant develop a timeline for tasks, understanding the time needed for each activity?
- **Problem-solving skills**: how does the participant handle unexpected challenges, demonstrating flexibility and adaptability in their planning?

Insight and compensatory strategies

Individuals with frontal lobe damage commonly exhibit insight deficits, lacking awareness of their cognitive impairments and their impact on daily functioning (Stuss and Benson, 1984; Scott and Schoenberg, 2011). This hinders their ability to employ effective compensatory strategies.

RLT Example: Ask the participant to reflect on a situation where they faced a cognitive challenge, such as managing multiple tasks simultaneously in a busy workplace environment, such as a restaurant kitchen or a retail store during a sale event and adapting to unexpected changes. The participant is instructed to imagine themselves in this scenario and describe how they would handle the situation. Inquire about their awareness of the difficulty and strategies employed to cope. Evaluate their ability to recognize and address cognitive impairments.

During the scenario, the clinician observes the participant's ability to recognize and address cognitive challenges in real-time. The participant may encounter unexpected changes or obstacles, such as a sudden influx of customers or equipment malfunctions. They are asked to verbalize their thoughts and actions as they navigate through the scenario, providing insights into their problem-solving strategies and coping mechanisms.

Key Aspects to Assess:

- **Awareness of cognitive challenges**: does the participant demonstrate an awareness of the cognitive demands of the scenario, such as the need to multitask and prioritize tasks effectively? Do they recognize the potential challenges they

may encounter, such as managing time constraints or dealing with unexpected events?

- **Employed strategies**: what strategies do the participant employ to cope with cognitive challenges and maintain performance? Do they demonstrate effective organization, time management, and decision-making skills in response to the demands of the scenario?
- **Flexibility and adaptability**: how does the participant respond to unexpected changes or disruptions in the scenario? Do they demonstrate flexibility and adaptability in adjusting their strategies and priorities to address new challenges as they arise?
- **Insight into cognitive impairments**: does the participant acknowledge any difficulties or limitations they experience during the scenario? Are they able to identify specific cognitive impairments or challenges they face, such as memory lapses or attention deficits?

Rule maintenance and cognitive flexibility

Frontal lobe damage can lead to impairments in rule maintenance and cognitive flexibility (Shallice and Burgess, 1996; Diamond, 2006). Individuals with such damage may struggle to maintain rules and adapt their behavior according to changing task demands, indicating deficits in executive functioning.

RLT Example: Modified "Uno" Rule Maintenance Task

Objective: Assess the participant's ability to maintain rules and adapt to changes in a modified version of the card game "Uno".

Instructions:

- Set up the game by shuffling the deck of Uno cards and dealing seven cards to each player, including the participant.
- Explain the basic rules of Uno and play several rounds.
- Then, change a rule or two in the game.
- Play several rounds of the modified Uno with the participant, ensuring they adhere to the rules and demonstrate understanding.
- Introduce variations and rule changes throughout the game to assess adaptability.
- Observe the participant's ability to maintain focus, follow evolving rules, and adapt strategy.
- Record any difficulties experienced in maintaining rules or adapting to modifications.

This modified Uno task provides a structured yet flexible assessment of rule maintenance and cognitive flexibility, mimicking real-life situations where individuals must adhere to rules and adjust their behavior accordingly.

Social cognition

Impairments in social cognition are frequently observed in individuals with frontal lobe damage (Knight and Grabowecky, 1995; Amodio and Frith, 2006). These individuals may struggle with interpreting social cues, understanding others' perspectives, and regulating their social behavior, reflecting deficits in social cognition.

RLT Example: Present a social scenario (e.g., a video clip or written description) and ask the participant to interpret emotions, intentions, and social dynamics.

An example can be a scene from the movie 'Forrest Gump', where Forrest attends a social gathering at his friend Lieutenant Dan's house. The subtext in this scene revolves around Forrest's innocence and straightforwardness contrasted with the complexity of social interactions happening around him.

Several aspects warrant attention:

- **Emotion interpretation**: assess the participant's understanding of the emotions experienced by the characters in the scenario. This involves identifying emotions accurately based on verbal and nonverbal cues. The clinician can ask the patients: What emotions do you think Forrest and the other characters are experiencing during the interaction?

- **Intention recognition**: evaluate the participant's ability to discern the intentions or motivations behind the words and actions of the characters. This involves inferring underlying motives from observable behaviors. The clinician can ask the patients: What do you believe are their intentions or motivations behind their words and actions?
- **Social dynamics**: analyze the participant's interpretation of the social dynamics between the characters. Determine whether they recognize the nature of the relationships, such as whether they are friendly, competitive, supportive, or indifferent. The clinician can ask the patients: How would you interpret the dynamics between Forrest and the other guests? Are they friendly, competitive, supportive, or indifferent?
- **Response to social cues**: consider how the participant would respond if they were in the situation depicted in the scenario. Assess their ability to appropriately react to social cues and interactions, taking into account their understanding of the context and their own social norms. The clinician can ask the patients: If you were Forrest in this situation, how would you respond to the various social cues and interactions?

Overall, attention should be given to the participant's comprehension of social nuances, their ability to accurately interpret social situations, and their capacity to respond appropriately to social cues, reflecting their social cognition abilities.

Discussion and conclusion

This opinion article aims to offer a broad trajectory for future explorations concerning the nuanced assessment of executive functioning deficits and their implications for RLT performance. Through the enhancement of assessment protocols and the inclusion of thorough observations encompassing RLTs of initiation, execution, organizational planning, social cognition, and insight—clinicians can acquire deeper insights into the functional capabilities of individuals with prefrontal cortex damage. Additionally, the classification of specific types of mistakes made

during task completion could inform targeted interventions tailored to address identified deficits. Overall, the integration of RLTs into neuropsychological evaluation holds promise for enhancing the accuracy, validity, and clinical utility of assessments for individuals with executive functioning impairments.

These RLT examples aim to evaluate various ecological dimensions of executive functioning associated with frontal lobe damage. However, it's essential to acknowledge that these examples represent only a subset of the challenges individuals with frontal lobe damage may face (Duncan, 1986; Delis et al., 2001; Stuss and Alexander, 2007; McCloskey et al., 2009; Damasio et al., 2011; Worthington, 2012; Otero and Barker, 2014). Additional domains, such as decision-making, can also be incorporated into RLT protocols, highlighting the need for ongoing development and refinement in this area.

It is crucial to customize tasks according to individual abilities and consider cultural and contextual factors during test administration.

This RLT approach might challenge the conventional practices of neuropsychologists, given that our professional training often centers on structured and standardized tests that yield normalized scores. However, in light of the "frontal lobe paradox", it becomes imperative to step outside the traditional framework and gather ecological data on the patient's actual executive abilities.

While concerns may arise regarding the lack of normative data for the proposed RLTs, their effectiveness is assessed based on success or failure, including partial success with specific types of errors, without relying on comparisons to established norms or standards. In other words, whether the tasks are successful or not can be judged based on their specific objectives and criteria, rather than comparing them to how others perform.

Presenting the results of these tasks, such as "X RLTs were administered to assess executive functioning, and patients failed to fulfill them successfully in a Y/X ratio... The type of errors/difficulties included…" can provide valuable additional insights in the report, particularly when complemented with etiology data and findings from brain imaging (MRI/CT), as well as outcomes from tailored tests to confirm or reject the presence of executive dysfunction.

In conclusion, the "frontal lobe paradox" presents a significant challenge in neuropsychological assessments, highlighting the need for a comprehensive approach that goes beyond traditional standardized tests. The incorporation of ecological validity assessment in the form of RLT, as proposed in this manuscript, may offers a potentially useful approach for addressing this paradox by providing a more nuanced understanding of individuals' cognitive functioning in real-world contexts.

Future research should focus on:

- Developing specific protocols for Real-Life Tasks (RLTs) and validating their effectiveness in assessing executive functioning deficits in individuals with frontal lobe damage.
- Empirically evaluating the efficacy of incorporating RLTs into neuropsychological evaluations for individuals with frontal lobe damage, comparing their outcomes with those of traditional standardized tests.
- Investigating the impact of ecological assessments, specifically RLTs, on treatment planning and outcomes for individuals with frontal lobe damage.
- Assessing the feasibility of implementing RLTs in routine clinical practice and evaluating their effectiveness in improving patient care and outcomes.

By systematically evaluating the benefits and limitations of incorporating RLTs, researchers and clinicians can better understand their role in addressing the real-world needs of individuals with frontal lobe damage. Ultimately, this approach can contribute to the refinement and optimization of neuropsychological assessment protocols, leading to improved assessment, care, and outcomes for this patient population.

This opinion manuscript serves as a call for further contemplation, research, and development in the context of integrating RLT into the standard neuropsychological assessment.

Author contributions

OE: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

Thanks to Zohar Rom for providing valuable feedback on the manuscript drafts, along with his insightful input.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Amodio, D., and Frith, C. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 7, 268–277. doi: 10.1038/nrn1884

Burgess, P.W., Alderman, N., Volle, E, Benoit, R.G., and Gilbert, S.J. (2009). Mesulam's frontal lobe mystery examined. (2009) Mesulam's frontal lobe mystery re-examined. *Restor. Neurol. Neurosci.*, 27, 493–506. doi: 10.3233/RNN-2009-0511

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, L. M., Dawson, D. R., et al. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *J. Int. Neuropsychol. Soc.: JINS.* 12, 194–209. doi: 10.1017/S1355617706060310

Damasio, A., Anderson, S. W., and Tranel, D. (2011). "The frontal lobes," in *Clinical Neuropsychology*, eds. K. M. Heilman and E. Valenstein (New York: Oxford University Press), 417–465.

Delis, D. C., Kaplan, E., and Kramer, J. (2001). *Delis-Kaplan Executive Function System*. San Antonio, TX: Psychological Corporation.

Diamond, A. (2006). "The early development of executive functions," in *Lifespan Cognition: Mechanisms of Change,* eds. E. Bialystok and F. I. M. Craik (New York: Oxford University Press).

Duncan, J. (1986). Disorganization of behaviour after frontal lobe damage. *Cogn. Neuropsychol.*, 2, 271–290. doi: 10.1080/02643298608253360

Fisher-Hicks, S., Wood, R. L., and QC, B. B. (2021). "The frontal lobe paradox," in *Neuropsychological Aspects of Brain Injury Litigation* (London: Routledge), 140–157.

George, M. S., and Gilbert, S. (2018). Mental Capacity Act 2005 assessments: why everyone needs to know about the frontal lobe paradox. *Neuropsychologist*. (2018) 5:59. doi: 10.53841/bpsneur.2018.1.5.59

Gioia, G., Isquith, P., Guy, S., and Kenworthy, L. (1996). *Behavior Rating Inventory of Executive Function*. Lutz, FL: Psychological Assessment Resources.

Goldstein, K., and Scheerer, M. (1941). Abstract and concrete behavior an experimental study with special tests. *Psychol. Monogr.* 53, i151. doi: 10.1037/h0093487

Goldstein, S., Naglieri, J. A., Princiotta, D., and Otero, T. M. (2013). "Introduction: A history of executive functioning," in *Handbook of Executive Functioning*, eds. S. Goldstein and J. A. Naglieri (New York, NY: Springer).

Knight, R. T., and Grabowecky, M. (1995). "Escape from linear time: Prefrontal cortex and conscious experience," in *The cognitive Neurosciences,* ed. M. S. Gazzaniga (Cambridge, MA: The MIT Press), 1357–1371.

Lezak, M. D., Howieson, D. B., Bigler, E. D., and Tranel, D. (2012). *Neuropsychological Assessment (5th ed.)*. Oxford: Oxford University Press.

Luria, A. R. (1980). "Disturbances of higher cortical functions with lesions of the frontal region," in *Higher Cortical Functions in Man* (Boston, MA: Springer).

McCloskey, G., Perkins, L. A., and Van Divner, B. R. (2009). *Assessment and Intervention for Executive Function Difficulties*. New York: Routledge.

Newstead, S., Lewis, J., Roderique-Davies, G., Heirene, R. M., and John, B. (2022). The paradox of the frontal lobe paradox. a scoping review. *Front. Psychiatry*, 13, 913230. doi: 10.3389/fpsyt.2022.913230

Otero, T. M., and Barker, L. A. (2014). "The frontal lobes and executive functioning," in *Handbook of Executive Functioning*, eds. S. Goldstein and J. A. Naglieri (Cham: Springer Science + Business Media), 29–44.

Scott, J. G., and Schoenberg, M. R. (2011). "Frontal lobe/executive functioning," in *The Little Black Book of Neuropsychology: A Syndrome-Based Approach*, eds. M. R. Schoenberg and J. G. Scott (Cham: Springer Science + Business Media), 219–248.

Shallice, T., and Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. Philosophical transactions of the Royal Society of London. *Series B, Biol. Sci.* 351, 1405–1412. doi: 10.1098/rstb.1996.0124

Stuss, D. T., and Alexander, M. P. (2007). Is there a dysexecutive system? *Philos. Trans. R. Soc. Lond,. B,. Biol. Sci.*, 362(1481), 901–915. doi: 10.1098/rstb.2007.2096

Stuss, D. T., and Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychol. Bull.* 95, 3–28. doi: 10.1037/0033-2909.95.1.3

Teuber, H.L. (1964). "The riddle of the frontal lobe function in man," in *The Frontal Granular Cortex and Behavior*, eds. J. M. Warren and K. Akert (New York: McGraw Hill), 410–458.

Walsh, K. W. (1985). *Understanding Brain Damage: A Primer of Neuropsychological Evaluation*. London: Longman Group Ltd.

Wood, L. I., and Bigler, E. (2017). "Problems assessing executive dysfunction in neurobehavioural disability," in *Neurobehavioural Disability and Social Handicap Following Traumatic Brain Injury*, eds. T. M McMillan and R. L. I Wood. (Oxford: Routledge), 88–100.

Worthington, A. (2019). Decision making and mental capacity: resolving the frontal paradox. *Neuropsychologist*. 7, 31–5. doi: 10.53841/bpsneur.2019.1.7.31

Worthington, A.D. (2012). "The natural recovery and treatment of executive disorders," in The Handbook of Clinical Neuropsychology, Second Edition, eds. J. M. Gurd, U. Kischka, and J. C. Marshall. (Oxford: Oxford University Press), 369–386.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 09 May 2024
DOI 10.3389/fpsyg.2024.1375326

OPEN ACCESS

EDITED BY
Elisa Cavicchiolo,
University of Rome Tor Vergata, Italy

REVIEWED BY
Laura Veronelli,
University of Milan-Bicocca, Italy
Lorenzo Diana,
IRCCS Istituto Auxologico Italiano, Italy

*CORRESPONDENCE
Yi Zhang
✉ zhangyizhe1975@163.com

RECEIVED 23 January 2024
ACCEPTED 25 April 2024
PUBLISHED 09 May 2024

CITATION
Shi Y and Zhang Y (2024) Reliability and validity of a novel attention assessment scale (broken ring enVision search test) in the Chinese population.
Front. Psychol. 15:1375326.
doi: 10.3389/fpsyg.2024.1375326

Reliability and validity of a novel attention assessment scale (broken ring enVision search test) in the Chinese population

Yue Shi and Yi Zhang*

Department of Rehabilitation Medicine, Third Affiliated Hospital of Soochow University, Changzhou, China

Background: The correct assessment of attentional function is the key to cognitive research. A new attention assessment scale, the Broken Ring enVision Search Test (BReViS), has not been validated in China. The purpose of this study was to assess the reliability and validity of the BReViS in the Chinese population.

Methods: From July to October 2023, 100 healthy residents of Changzhou were selected and subjected to the BReViS, Digital Cancelation Test (D-CAT), Symbol Digit Modalities Test (SDMT), and Digit Span Test (DST). Thirty individuals were randomly chosen to undergo the BReViS twice for test–retest reliability assessment. Correlation analysis was conducted between age, education level, gender, and various BReViS sub-tests including Selective Attention (SA), Orientation of Attention (OA), Focal Attention (FA), and Total Errors (Err). Intergroup comparisons and multiple linear regression analyses were performed. Additionally, correlation analyses between the BReViS sub-tests and with other attention tests were also analyzed.

Results: The correlation coefficients of the BReViS sub-tests (except for FA) between the two tests were greater than 0.600 ($p < 0.001$), indicating good test–retest reliability. The Cronbach's alpha coefficient was 0.874, suggesting high internal consistency reliability. SA showed a significant negative correlation with the net score of D-CAT ($r = -0.405$, $p < 0.001$), and a significant positive correlation with the error rate of D-CAT ($r = 0.401$, $p < 0.001$), demonstrating good criterion-related validity. The correlation analysis among the results of each sub-test showed that the correlation coefficient between SA and Err was 0.532 ($p < 0.001$), and between OA and Err was-0.229 ($p < 0.05$), whereas there was no significant correlation between SA, OA, and FA, which indicated that the scale had good informational content validity and structural validity. Both SA and Err were significantly correlated with age and years of education, while gender was significantly correlated with OA and Err. Multiple linear regression suggested that Err was mainly affected by age and gender. There were significant differences in the above indexes among different age, education level and gender groups. Correlation analysis with other attention tests revealed that SA negatively correlated with DST forward and backward scores and SDMT scores. Err positively correlated with D-CAT net scores and negatively with D-CAT error rate, DST forward and backward scores, and SDMT scores. OA and FA showed no significant correlation with other attention tests.

Conclusion: The BReViS test, demonstrating good reliability and validity, assessing not only selective attention but also gauging capacities in immediate memory, information processing speed, visual scanning, and hand-eye coordination. The results are susceptible to demographic variables such as age, gender, and education level.

KEYWORDS

attention, attention assessment, broken ring enVision search test, reliability, validity, age, education level, gender

1 Introduction

Attention is the foundation of all cognitive functions, the prerequisite for continuous information processing, and a gateway for the flow of information to enter the brain and undergo selection (Petersen and Posner, 2012). Precise and accurate assessment of attentional functions is key in cognitive research and a precondition for the rehabilitation of cognitive disorders. In clinical neuropsychology, visual search tasks (VSTs) are frequently used to evaluate selective visual attention deficits in patients with neurological conditions (Eglin et al., 1989; Luck et al., 1989; Utz et al., 2013). These typically include paper-and-pencil target cancellation tasks such as the Attention Matrix (Della Sala et al., 1992), Ruff 2&7 Selective Attention Test (Marioni et al., 2012), Letter Cancellation Test (Uttl and Pilkenton-Taylor, 2001), and the Visual Spatial Attention subtest in the Oxford Cognitive Screen (Demeyere et al., 2015), which are effective tools for detecting attention deficits post-stroke. However, existing VSTs do not take into account the potential impact of stimulus layout and crowding on the test results of participants. Facchin et al. developed a novel attention assessment scale—the Broken Ring enVision Search Test (BReViS) to evaluate attentional functions (Facchin et al., 2023). It assesses different components of attention including selective attention, the visual–spatial orientation of attention, and focal attention involving crowding phenomena, and is a novel open-ended paper-and-pencil assessment tool.

While studies have shown the effectiveness and applicability of the BReViS test in the Italian population and provided specific Italian normative data, its suitability for the Mainland Chinese population is yet to be concluded. Therefore, this study aims to examine the reliability and validity of the BReViS test in the healthy Chinese population and to analyze the characteristics of its preliminary application, in the hope of finding a simple and feasible tool for the clinical environment to assess neuropsychological patients' attention deficits and provide a basis for the assessment and rehabilitation treatment of attentional disorders.

2 Sample and methods

2.1 Study procedure

General Information: From July to October 2023, a total of 100 healthy residents, including staff and accompanying personnel from the First People's Hospital of Changzhou and residents of Tianning and Xinbei districts of Changzhou, were selected. The cohort comprised 47 males and 53 females; ages ranged from 19 to 84 years, with an average age of (52.35 ± 22.01) years; years of education ranged from 2 to 20 years, with an average of (12.39 ± 3.86) years. Of these, the number of people with 2 years of education was 1.

Inclusion criteria: Age 19–84 years; Right-handed; Normal or corrected-to-normal vision.

Exclusion criteria: Auditory, visual, or speech impairments; Past history of neurological or psychiatric diseases (including brain injury, stroke, clinically diagnosed dementia, depression, etc.); History of addiction to tobacco, alcohol, or addictive drugs.

Grouping method: In order to make between-group comparisons between different ages, education levels and genders, the subjects were divided into 4 groups according to different ages in the statistical analyses, with those aged 18–34 years classified as the youth group, those aged 35–49 years classified as the young-adult group, those aged 50–65 years classified as the middle-aged group, and those older than 65 years classified as the senior group. Similarly, they were divided into four groups according to their education level: those with 1–6 years of education were classified as the elementary group, those with 7–9 years of education were classified as the middle school group, those with 10–12 years of education were classified as the high school/vocational group, and those with more than 12 years of education were classified as the college/university and above group. They were divided into male and female groups by gender. Demographic characteristics of the groups are reported in Table 1. Thirty subjects were randomly selected as the retesting group and the BReViS test was administered again after 2 weeks. There were 30 subjects in the retesting group, of whom 14 were male and 16 were female; their ages ranged from 19 to 72 years, with a mean of (44.07 ± 15.67) years; and their years of education ranged from 6 to 19 years, with a mean of (13.86 ± 2.81) years.

2.2 Measurements and applied questionnaires

2.2.1 The BReViS test

It was developed by Facchin et al. (2023). We have obtained authorization from the original authors to use it. The test consisted of four cancellation quiz cards, each consisting of five rows of circles with notches in different orientations arranged in different layouts and degrees of crowding, with 25 targets per card and randomly defined target locations. Subjects were asked to identify and cross out all the targets on each card that had the same notch orientation as the circles shown at the top of the card, and to record the execution time, number of omissions, self-corrections, and errors crossings for the completion

TABLE 1 Demographic characteristics of the patients' sample.

Age	19–34		35–49		50–65		>65		Tot.
School	F	M	F	M	F	M	F	M	
1–6	0	0	0	0	1	0	5	4	10
7–9	0	0	0	1	2	2	7	6	18
10–12	0	1	0	1	1	2	11	8	24
>12	20	15	4	3	1	0	1	4	48
Tot.	20	16	4	5	5	4	24	22	100

F = female; M = male.

of the 4 test cards. The performance time for each quiz card was calculated based on the execution time and omissions for each card. The calculation formula is as follows:

$$\text{Performance time} = \frac{25 \times \text{Execution time}}{25 - \text{omissions}}$$

By combining the execution times of the four test cards, the following four indices are calculated: Selective Attention (SA), Orientation of Attention (OA), Focal Attention (FA), and Total Errors (Err).

SA represents the capacity to suppress irrelevant stimuli (distractors) and solely select relevant stimuli (targets) under the simplest conditions. It directly corresponds to the performance time of the first card (linear layout, low crowding), which is less affected by random arrays and crowded displays. SA = Performance time for the first card. Higher SA index values suggest lower efficiency of selective attention.

OA refers to the strategic direction of visual attention, which is the capacity to guide selective visual attention with effective endogenous strategies throughout the visual scene (Connor et al., 2004), one of the two components of visual–spatial attention measured by BReViS. High OA index values indicate an inability to follow effective endogenous strategies during the visual search process, necessitating exogenous cues to perform the task correctly. It is calculated with the following formula using the performance time of each card:

$$\text{OA} = \frac{\text{Card3} + \text{Card4}}{2} - \frac{Card1 + Card2}{2}$$

FA can be interpreted as the ability to adjust the focus of attention based on the position of stimuli within the array, another component of visual–spatial attention (Castiello and Umilta, 1990). It corresponds to the comparison between two levels of crowding: high and low. High FA index values suggest a higher sensitivity to crowding. It is calculated with the following formula using the performance time of each card:

$$\text{FA} = \frac{\text{Card2} + \text{Card4}}{2} - \frac{Card1 + Card3}{2}$$

The Err index represents the overall errors made across all sub-tests. Err = Total number of errors across all four test cards.

2.2.2 Other attention tests

The Digit Cancellation Test (D-CAT) is used to measure selective attention (Hatta et al., 2004). Participants were required to locate and strike through the number preceding the number 3 from a random sequence of numbers 1–9, with the time taken to complete the test recorded. Net scores and error rates are calculated based on the number of correct cancelations, omissions, and mistakes. Higher net scores and lower error rates indicate better selective attention.

The Symbol Digit Modalities Test (SDMT) was published by Aaron Smith in 1973 and revised in 1982 to assess speed of information processing, visual scanning ability, and hand-eye coordination (Strober et al., 2019). This test involves an encoding key of 9 different abstract symbols, each associated with a number. Participants must write the number corresponding to each symbol as quickly as possible within 90 s. Scoring is based on the number of correct symbols and reversed symbols. Higher scores indicate better speed of information processing, visual scanning ability, and hand-eye coordination.

The Digit Span Test (DST) is a commonly used psychological assessment tool that measures short-term memory and attention span (Park and Lee, 2019). In its traditional form, the Digit Span Test consists of two parts: forward digit span and backward digit span. This test evaluates the participant's ability to recall a sequence of numbers in the correct order both forwards and backwards after the tester reads them out. Participants repeat a series of random numbers at a rate of one number per second, starting with a sequence of 3 numbers and increasing in length up to 12 numbers or until two consecutive errors are made. One point is scored for each correctly recalled sequence. The higher the scores on forward and backward digit span, the greater the capacity of immediate memory.

2.2.3 Sample size calculation

This study mainly used correlation analysis and multiple linear regression analysis, so it was calculated using G*Power software 3.1 (Faul et al., 2009), correlation analysis input target effect size of 0.3, type I error of 5% ($\alpha = 0.05$), and power of 80% ($\beta = 0.20$), and the sample size of 82 participants was calculated. Multiple linear regression analyses were conducted with an input independent variable of 3 (U = 3), effect size = 0.15 (F2 = 0.15), type I error of 5% ($\alpha = 0.05$), and power of 80% ($\beta = 0.20$), resulting in a calculated sample size of 77 participants. The final sample size was 100 participants, taking into account an allowable 20% dropout rate.

2.2.4 Experimental procedure

Participants filled out informed consent forms; They were subjected to the BReViS test and other attention tests. Among them, 30 were randomly selected to retake the BReViS test after two weeks. All tests were administered by the same physician.

2.3 Statistical analysis

SPSS 17.0 software was used for statistical analysis. Spearman's correlation analysis was employed to assess the correlation between the BReVis test and other attention tests, as well as the correlation between each sub-test of the BReViS and age, educational level, and gender. Kruskal-Wallis test was used to compare the differences in the BReViS sub-test scores among different age and educational level groups, while Mann–Whitney U test was utilized to compare the differences between gender groups. Multiple linear regression analysis was conducted to investigate the influence of demographic characteristics on scale evaluation results, with statistical significance set at $p < 0.05$. Pearson correlation coefficient was employed to analyze the test–retest reliability of the BReViS; Cronbach's α coefficient was used to indicate internal consistency, with a coefficient above 0.80 considered excellent, between 0.70 and 0.80 acceptable, and below 0.7 indicating poor reliability. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were employed to analyze the appropriateness of factor analysis, to validate the structural validity of the BReViS. Finally, correlation analyses between the results of the BReViS subtests were conducted using Spearman's correlation analysis to test the content and structural validity of the scale.

TABLE 2 Mean performance time (and SD) for each sub-test, divided by age group.

Sub-test	19–34	35–49	50–65	>65
SA	49.95 (12.36)	64.28 (14.37)	79.58 (18.56)	98.73 (27.77)
OA	30.06 (18.83)	30.44 (17.93)	26.50 (27.28)	28.89 (36.01)
FA	−0.69 (12.10)	−5.56 (10.14)	1.83 (13.72)	4.09 (17.52)
Err	9.33 (5.78)	13.56 (7.27)	19.89 (8.34)	22.41 (10.21)

TABLE 3 Mean performance time (and SD) for each sub-test, divided by education level group.

Sub-test	1–6	7–9	10–12	>12
SA	101.63 (22.02)	104.3 (26.86)	86.72 (26.76)	55.41 (18.90)
OA	39.05 (32.33)	17.28 (40.84)	36.88 (28.02)	27.85 (19.87)
FA	0.55 (18.44)	8.28 (17.33)	0.42 (17.16)	−0.73 (11.45)
Err	27.3 (11.37)	20.22 (12.07)	18.88 (7.96)	12.04 (7.82)

3 Results

3.1 Descriptive results

The descriptive mean results on the four BReViS sub-tests scores are reported in Tables 2–4.

TABLE 4 Mean performance time (and SD) for each sub-test, divided by gender group.

Sub-Test	Male	Female
SA	76.66 (26.55)	76.07 (34.40)
OA	36.30 (26.78)	22.97 (28.42)
FA	−0.15 (14.71)	2.58 (15.23)
Err	14.06 (7.62)	19.00 (11.77)

3.2 Correlation analysis of age with the BReViS sub-tests

Age showed a positive correlation with both SA ($r = 0.776$, $p < 0.001$) and Err ($r = 0.607$, $p < 0.001$), with no significant correlation with the other sub-tests.

3.3 Comparison of different age groups

As shown in Table 5, analyses of multiple between-group comparisons across age groups showed significant differences in sub-test scores for SA and Err ($p < 0.001$). Detailed two-by-two intergroup comparisons highlighted significant differences in SA scores between the youth and middle-aged groups (adjusted $p = 0.006$), as well as between the youth and senior groups (adjusted $p = 0.000$). Similarly, Err scores differed significantly between the young and middle-aged groups (adjusted $p = 0.005$), and between the youth and senior groups (adjusted p = 0.000). Additionally, a distinct variance was observed in SA scores between the young-adult and senior groups (adjusted $p = 0.017$), as shown in Table 6.

TABLE 5 Analysis of variance between different age groups (Mean Rank).

Sub-test	Youth group	Young-adult group	Middle-aged group	Senior group	*p*
SA	22.53	41.11	58.22	72.72	0.000
OA	51.81	52.11	44.67	50.30	0.926
FA	47.72	36.22	53.44	54.89	0.301
Err	27.25	42.33	63.56	67.74	0.000

3.4 Correlation analysis of education level with the BReViS sub-tests

Years of education were negatively correlated with both SA ($r = -0.715$, $p < 0.001$) and Err ($r = -0.502$, $p < 0.001$), with no significant correlation with the remaining sub-tests.

3.5 Comparison of different education level groups

As shown in Table 7, analyses of multiple between-group comparisons across education level groups unveiled significant disparities in the scores for sub-tests SA and Err, while OA and FA did not exhibit such differences ($p < 0.001$). Detailed two-by-two intergroup comparisons highlighted significant differences in SA scores: the college/university and above group demonstrated significant disparities when compared with the elementary, middle school, and high school/vocational groups (adjusted $p = 0.000$ for all comparisons). Similarly, Err scores significantly differed between the college/university and above group and the elementary group (adjusted p = 0.000), as well as between the college/university and above group and both the middle school (adjusted $p = 0.027$) and high school/vocational groups (adjusted $p = 0.006$), as detailed in Table 8.

3.6 Correlation analysis of gender with the BReViS sub-tests

Gender showed a negative correlation with OA ($r = -0.251$, $p = 0.012$) and a positive correlation with Err ($r = 0.215$, $p = 0.032$), with no significant correlation with SA and FA.

3.7 Comparison of the two gender groups

The comparison results between the two gender groups showed a significant difference in OA and Err ($p < 0.05$), while no significant difference was observed in SA and FA, as detailed in Table 9. Combining the results from Table 4, it was evident that males scored higher in the OA test and lower in the Err test compared to females.

TABLE 6 Two-by-two comparison of SA and Err between different age groups.

Sample 1-Sample 2	SA			Err		
	Test Statistic	S.E	Adj.*p*	Test Statistic	S.E	Adj.*p*
1–2	−18.58	10.81	0.514	−15.08	10.80	0.976
1–3	−35.69	10.81	0.006	−36.31	10.80	0.005
1–4	−50.19	6.46	0.000	−40.49	6.45	0.000
2–3	−17.11	13.68	1.000	−21.22	13.67	0.723
2–4	−31.61	10.57	0.017	−25.41	10.57	0.097
3–4	−14.50	10.57	1.000	−4.18	10.57	1.000

1 = the youth group; 2 = the young-adult group; 3 = the middle-aged group; 4 = the senior group.

TABLE 7 Analysis of variance between different education level groups (mean rank).

Sub-test	Elementary group	Middle school group	High school/ vocational group	College/University group and above	*p*
SA	77.05	76.97	62.73	28.93	0.000
OA	56.90	41.14	57.54	49.16	0.275
FA	49.10	61.17	50.33	46.88	0.361
Err	79.60	59.06	60.29	36.33	0.000

TABLE 8 Two-by-two comparison of SA and Err between different education level groups.

Sample 1-Sample 2	SA			Err		
	Test Statistic	S.E	Adj.*p*	Test Statistic	S.E	Adj.*p*
4–3	33.80	7.25	0.000	23.96	7.25	0.006
4–2	48.05	8.02	0.000	22.72	8.01	0.027
4–1	48.12	10.08	0.000	43.27	10.08	0.000
3–2	14.24	9.05	0.692	1.236	9.04	1.000
3–1	14.32	10.92	1.000	19.31	10.91	0.461
2–1	0.08	11.44	1.000	20.54	11.43	0.434

1 = the elementary group; 2 = the middle School group; 3 = the high school/vocational group; 4 = the college/university group and above.

TABLE 9 Comparison of the two gender groups (Mean Rank).

Sub-test	Male	Female	*p*
SA	52.61	48.63	0.494
OA	58.19	43.68	0.013
FA	46.07	54.42	0.151
Err	43.91	56.34	0.032

3.8 Impact of demographic variables

Multiple linear regression analysis suggested that when demographic variables age, education level, and gender were introduced into the linear regression model of SA and Err, SA was affected by years of education level and age, while Err was influenced by age and gender (Table 10).

3.9 Relevance to other attention tests

SA was negatively correlated with the net score of D-CAT and positively correlated with the error rate of D-CAT. It was also negatively correlated with DST forward and backward scores and SDMT scores. Err showed a positive correlation with the net score of D-CAT and a negative correlation with the error rate of D-CAT, DST forward and backward scores, and SDMT scores. OA and FA did not show significant correlation with other attention tests (Table 11).

3.10 Reliability testing

3.10.1 Re-testability of the BReViS test: Results showed that the correlation coefficients for SA, OA, and Err were all greater than 0.600, $p < 0.001$. Only the correlation coefficient for FA was below 0.6, $p > 0.05$, which was not statistically significant (Table 12).

3.10.2 Internal Consistency Reliability: Cronbach's alpha coefficient was 0.874, indicating high internal consistency reliability for the BReViS test.

3.11 Validity testing

3.11.1 Construct Validity: The Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity results were 0.763 and 252.601

(P<0.001), respectively, indicating the scale was not very suitable for factor analysis.

3.11.2 Criterion Validity: In this study, the D-CAT was used as a criterion, and Spearman's correlation analysis was used to calculate the correlation between BReViS's SA and the net scores and error rates of D-CAT to evaluate the degree of criterion-related validity. The results showed that SA was significantly negatively correlated with the net score of D-CAT ($r = -0.405$, $p < 0.001$) and significantly positively correlated with the error rate of D-CAT ($r = 0.401$, $p < 0.001$), indicating the questionnaire has good criterion-related validity, as seen in Table 11.

3.12 Correlation between sub-tests

The correlation analysis of the results among the various sub-tests of the BReViS test indicated that the correlation coefficient between SA and Err was 0.532, and between OA and Err was-0.229, with $p < 0.05$, suggesting a certain degree of consistency between them, which contributes to ensuring the reliability of the scale. Meanwhile, the correlation between SA, OA, and FA was not high, indicating that the scale has excellent information content and structural validity, as seen in Table 13.

4 Discussion

Attention is a fundamental psychological concept, deeply embedded in cognitive processing, defined by the deliberate focusing on particular stimuli (van Es et al., 2018). This focusing elevates the level of awareness about these stimuli, epitomizing attention's selective nature. Solso, MacLin M.K., and MacLin O.H. (2005) highlight that "the essence of attention lies in the concentration and focus of consciousness," underlining attention's critical role in selecting an item from an array of simultaneous stimuli or thought sequences (Baddeley, 1988). Selective attention, therefore, is the capacity to direct an individual's finite processing resources toward a particular environmental aspect. This complex concept encompasses a range of processes, including spatial attention with its directional and focal elements (Carrasco, 2011). Such capability allows for the filtration of extensive information from the surroundings, facilitating the efficient usage of scarce cognitive resources.

Historically, attention has been a central theme in psychological studies, resulting in a plethora of theoretical frameworks and experimental methodologies. One of the most significant paradigms for investigating selective visual attention's traits is visual search (Bacon and Egeth, 1997; Verghese, 2001; Wolfe, 2003). Everyday life is replete with visual search scenarios, whether it's choosing products on supermarket shelves, animals searching for food amidst leaves, locating a friend in a large gathering, or playing visual search games (Wolfe, 2020). Clinical neuropsychology frequently employs visual search tasks (VST) to evaluate selective visual attention deficits in patients with neurological conditions (Senger et al., 2017). Standard VST

TABLE 10 Impact of demographic variables.

Scale		B	S. E	t	p
Err	Age	0.281	0.036	7.728	0.000
	Gender	5.855	1.594	3.673	0.000
	Education level	−0.489	0.263	−1.863	0.066
	Constant	−6.977	3.300	−2.114	0.037
SA	Age	0.803	0.117	6.844	0.000
	Gender	1.753	4.053	0.432	0.666
	Education level	−2.088	0.668	−3.125	0.002
	Constant	60.167	13.194	4.560	0.000

TABLE 12 Re-testability of the BReViS test.

Index	Mean performance time (and SD) for the first test	Mean performance time (and SD) for the second test	r	p
SA	54.73 (15.10)	53.79 (14.27)	0.782	0.000
OA	29.60 (19.50)	22.17 (14.53)	0.659	0.000
FA	−1.57 (12.33)	0.67 (9.67)	0.110	0.564
Err	9.87 (6.17)	10.10 (5.55)	0.759	0.000

TABLE 13 Correlation between sub-tests.

Index	r	p
SA and OA	−0.004	0.971
SA and FA	0.074	0.462
SA and Err	0.532	0.000
OA and FA	−0.050	0.621
OA and Err	−0.229	0.022
FA and Err	−0.012	0.904

TABLE 11 Relevance to other attention tests.

	SA		OA		FA		Err	
	r	p	r	p	r	p	r	p
D-CAT net score	−0.405	0.000	0.046	0.648	−0.045	0.658	−0.439	0.000
D-CAT error rate %	0.401	0.000	−0.048	0.635	−0.044	0.660	0.437	0.000
DST forward score	−0.624	0.000	0.035	0.732	−0.170	0.091	−0.458	0.000
DST backward score	−0.643	0.000	−0.046	0.646	−0.171	0.089	−0.417	0.000
SDMT score	−0.802	0.000	−0.059	0.557	−0.155	0.124	−0.529	0.000

protocols involve participants identifying a target among numerous stimuli, like figures or letters, assessing performance based on response accuracy and time (Wolfe et al., 2002).

Studies suggest that visual task outcomes are not just influenced by attention toward the target's location (the spatial component) but also by adjusting the attention window according to the task requirements (the focal component) (Albonico et al., 2016), with each component operating independently (Castiello and Umilta, 1990; Carrasco and Yeshurun, 2009). Traditional VSTs, however, tend to neglect the influence of distractor arrangement and density on performance, thus failing to adequately capture the nuances of spatial attention (Weintraub and Mesulam, 1988; Mesulam, 2000). The BReViS assessment offers a refreshing alternative to conventional paper-and-pencil visual search tests by modifying the stimulus arrangement within the visual field, allowing for a comprehensive evaluation of selective visual attention and its distinct facets. Though previously utilized within the Italian demographic without undergoing thorough reliability and validity verification, this study introduces the BReViS test to the Mainland Chinese audience, undertaking a comprehensive examination of its reliability and validity among individuals aged 19 to 84.

4.1 Reliability testing

When a test has good reliability, it will yield almost the same scores for the same group of people at different times. The quality of reliability is also a prerequisite for validity testing. In this study, the test–retest reliability of the BReViS showed high correlation coefficients for three of the four sub-tests—SA, OA, and Err—on reassessment after two weeks. The test–retest results indicate that the BReViS test has good retest reliability, suggesting good temporal stability. The lack of statistical significance for FA in the correlation analysis may be due to the longer duration of this test, which may lead to fatigue in older participants resulting in unstable scores. Additionally, a higher Cronbach's alpha coefficient indicates stronger internal consistency of the scale. It is generally considered that a Cronbach's alpha coefficient greater than 0.7 indicates good consistency among items (Tavakol and Dennick, 2011). The results of this study show a total Cronbach's alpha coefficient of 0.874 for the BReViS test, indicating high internal consistency reliability. It's interesting to note that the average score for FA increased from −1.57 in the first test to 0.67 in the second, indicating a higher sensitivity to crowding in the latter. Research has shown that sensitivity to visual crowding is influenced by various factors that can affect an individual's ability to distinguish objects in cluttered environments. These factors include contrast, eccentricity, visual acuity and age, spatial frequency, attention and perceptual learning, as well as stimulus similarity (Coates et al., 2013; Veríssimo et al., 2022). Therefore, factors such as the brightness of the room, the depth of color of the test figures, the position of the test paper in the field of vision, whether the participant is focused, has undergone perceptual learning, and the objects surrounding the test paper can all affect sensitivity to crowding. The variability in the results of the two tests in this study reminds us that these influences need to be more tightly controlled in future studies.

4.2 Validity testing

The Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test suggested that the structure of the BReViS test might not be well suited for factor analyses, but that there was some correlation between the BReViS measures. The correlation analysis among the results of each subt-est of the BReViS showed a correlation coefficient of 0.532 between SA and Err, and −0.229 between OA and Err, with $p < 0.05$, indicating a certain level of consistency between them, which contributes to ensuring the reliability of the scale. However, the correlations among SA, OA, and FA were not high, suggesting that the scale has excellent information content and structural validity. Given that BReViS was developed to assess SA, this study employed the D-CAT as a criterion measure and found a significant correlation between SA and the D-CAT results, indicating good criterion-related validity.

4.3 The influence of age on BReViS

This study showed that age was significantly positively correlated with the sub-tests SA and Err. Multiple linear regression analysis suggested that SA is greatly influenced by age and education level, while Err is more influenced by age and gender. Therefore, age is a major factor influencing BReViS test results, which is consistent with the findings of the scale developers in the Italian population and previous research. The rank-sum test analysis across different age groups reveals that young adults significantly outperform both middle-aged and senior groups in selective attention tasks, making fewer errors. Additionally, the young-adult group demonstrate superior selective attention capabilities compared to those in the senior group. This pattern supports the notion that selective attention abilities undergo a pronounced growth during adolescence, which is then followed by a discernible decline as individuals age (Moore and Zirnsak, 2017). Neurophysiological alterations, observable through changes in the amplitude and latency of event-related potential (ERP) components, accompany this evolution in attention processing (Madden et al., 2007). Complementing these findings, functional MRI studies have identified a diminished activation in critical regions associated with visual attention control - namely, the bilateral fusiform gyrus, the right lingual gyrus, and the right precuneus-in elderly individuals when compared to their younger counterparts (Lyketsos et al., 1999; Lee et al., 2003).

4.4 The influence of education level on BReViS

This study found that years of education were negatively correlated with both SA and Err, and significant differences in SA and Err scores were also observed across different education level groups. Analysis using rank-sum tests across different educational attainment groups indicates that individuals with tertiary education (the college/university group and above) perform significantly better in selective attention tasks than those from the elementary (Mueller et al., 2008; Yehezkel et al., 2015), middle School and high school/vocational groups. They made fewer errors, suggesting a correlation between higher education levels and improved selective attention abilities.

Studies have shown that individuals with higher levels of education often perform better on various cognitive tests (Lindenberger and Baltes, 1997; Hultsch et al., 1999), likely due to the enhanced cognitive strategies, problem-solving skills, and knowledge base provided by formal education. Additionally, higher education may mitigate the impact of aging on cognitive performance (Lee et al., 2003; Jones et al., 2006; Tun and Lachman, 2008; Marioni et al., 2012). Research by Stern et al. (2005) and others indicates that higher educational attainment can moderate the decline in reaction and attention abilities due to aging and lower the risk of dementia (Bell et al., 2006), partly because cognitive reserve accumulation improves brain network efficiency (Rubia et al., 2010). These findings highlight the importance of considering educational background when interpreting cognitive assessment results.

4.5 The influence of gender on BReViS

In this study, the SA index was influenced by age and educational level, but no significant gender differences were observed. Gender was positively correlated with the Err index and negatively correlated with the OA index, with significant differences between genders, indicating that females committed more total errors than males. Males had higher OA scores than females, suggesting that males in the visual search process rely on exogenous cues to perform tasks correctly and are less likely to follow effective endogenous strategies. This is consistent with the observations made by the authors in a normal Italian population. The differences in OA scores between males and females may be related to the activation of different brain regions during the execution of spatial selective attention tasks. Males show increased activation in the left hemisphere's inferior parietal lobule, while females show significant activation in the right hemisphere's inferior frontal gyrus, insula, caudate, and temporal areas (de Fockert et al., 2001; Boi et al., 2011), which may be related to the modulation by estrogen and testosterone (Oberauer, 2019). Additionally, FA was not observed to be affected by gender, age and years of education in this study, which is in line with the results of the most recent application of the scale, i.e., crowding did not worsen with age (Pegoraro et al., 2024), and these findings are consistent with previous studies (Malavita et al., 2017; Shamsi et al., 2022).

4.6 The correlation between BReViS and other attention scales

SA was significantly positively correlated with the cancellation time and error rate in the D-CAT and significantly negatively correlated with the net score of cancellation. Err was negatively correlated with the net score of cancellation and positively correlated with the cancellation error rate. These results indicate that BReViS's SA and Err have good consistency with the D-CAT in assessing selective attention in the normal population.

Research demonstrates that enhancing selective attention significantly improves test outcomes in immediate memory capabilities (Plebanek and Sloutsky, 2019). For instance, within the context of the DST, superior selective attention enables individuals to recall and reproduce digit sequences with greater accuracy, thus exhibiting an increased memory capacity. This study reveals a negative correlation between SA and Err with the scores of forward and backward span in the DST, offering a crucial insight: higher scores of SA and Err indicate weaker selective attention, an increased error rate, and a noticeable decline in the subjects' immediate memory capacity. This finding highlights the close interrelation among immediate memory, selective attention, and cognitive efficiency, suggesting that individuals with a larger immediate memory capacity can more effectively resist distractions, thereby reducing error rates (Posner and Petersen, 1990; Rayner, 1998; Ku, 2018). In clinical practice, this correlation is important to identify and assess deficits in attention, working memory, or other cognitive functions.

The negative correlation between SA and Err with scores on the SDMT unveils a significant cognitive phenomenon: there is a direct correlation between elevated selective attention and increased efficiency of visual scanning, speed of information processing, and hand-eye coordination. Selective attention, a critical dimension of attention management, involves filtering task-relevant information from the environment while disregarding irrelevant distractions (De la Torre et al., 2015). The efficacy of selective attention depends to a large extent on the efficiency of visual scanning, a crucial aspect because it requires the individual to quickly localize and identify key targets among numerous visual stimuli (Reigal et al., 2019). Furthermore, the acceleration of information processing speed is a key factor in enhancing the efficiency of selective attention, allowing individuals to recognize important information within shorter durations and respond accordingly (Posner, 1980). In tasks requiring rapid identification of visual information followed by corresponding physical actions, exceptional hand-eye coordination markedly improves the precision and efficiency of task execution (Castiello and Umilta, 1990). Thus, the effective concentration of selective attention on specific stimuli or tasks is supported by an individual's performance in terms of a combination of speed of information processing, visual scanning ability, and hand-eye coordination. The improvement of these cognitive abilities not only further enhances the performance of selective attention but also, reciprocally, enhances the operational efficacy of these cognitive functions, thereby creating a positive feedback loop. This phenomenon offers profound insights into how individuals process information efficiently in complex environments within the domain of cognitive science.

The allocation of attentional resources in space involves two distinct processes: the orienting process, which selectively concentrates on specific aspects of the environment while ignoring others. The OA index reflects orienting ability, influenced by factors like stimulus salience, personal interests or goals, and the presence of attention-directing cues (Chun et al., 2011). The focusing process narrows attention to a specific area or object, acting like a magnifying glass, allowing selective concentration on a limited spatial area (Turatto et al., 2000; Chun et al., 2011). The FA index reflects focusing ability. Some studies suggest that focusing and orienting may vary based on visual conditions (Turatto et al., 2000). This research found no significant correlation between OA and FA with DST and SDMT, suggesting that orienting and focusing abilities might not be affected by immediate memory capacity, information processing speed, visual scanning ability, and hand-eye coordination skills.

5 Conclusion

The BreViS test, demonstrating good reliability and validity, is adept for application across a broad age range (19 to 84 years) within the general population, assessing not only selective attention but also gauging capacities in immediate memory, information processing speed, visual scanning, and hand-eye coordination. The influence of demographic variables such as age, gender, and education level on test outcomes underscores the necessity for nuanced interpretation of results in research and clinical settings.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the Third Affiliated Hospital of Soochow University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YS: Writing – original draft. YZ: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albonico, A., Malaspina, M., Bricolo, E., Martelli, M., and Daini, R. (2016). Temporal dissociation between the focal and orientation components of spatial attention in central and peripheral vision. *Acta Psychologica* 171, 85–92. doi: 10.1016/j.actpsy.2016.10.003

Bacon, W. J., and Egeth, H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 948–961. doi: 10.1037/0096-1523.23.4.948

Baddeley, A. (1988). Cognitive psychology and human memory. *Trends Neurosci.* 11, 176–181. doi: 10.1016/0166-2236(88)90145-2

Bell, E. C., Willson, M. C., Wilman, A. H., Dave, S., and Silverstone, P. H. (2006). Males and females differ in brain activation during cognitive tasks. *NeuroImage* 30, 529–538. doi: 10.1016/j.neuroimage.2005.09.049

Boi, M., Vergeer, M., Ogmen, H., and Herzog, M. H. (2011). Nonretinotopic exogenous attention. *Curr. Biol.* 21, 1732–1737. doi: 10.1016/j.cub.2011.08.059

Carrasco, M. (2011). Visual attention: the past 25 years. *Vis. Res.* 51, 1484–1525. doi: 10.1016/j.visres.2011.04.012

Carrasco, M., and Yeshurun, Y. (2009). Covert attention effects on spatial resolution. *Prog. Brain Res.* 176, 65–86. doi: 10.1016/S0079-6123(09)17605-7

Castiello, U., and Umilta, C. (1990). Size of the attentional focus and efficiency of processing. *Acta Psychol.* 73, 195–209. doi: 10.1016/0001-6918(90)90022-8

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annu. Rev. Psychol.* 62, 73–101. doi: 10.1146/annurev.psych.093008.100427

Coates, D. R., Chin, J. M., and Chung, S. T. (2013). Factors affecting crowded acuity: eccentricity and contrast Optometry and vision science. *Am. Acad. Optom.* 90, 628–638. doi: 10.1097/OPX.0b013e31829908a4

Connor, C. E., Egeth, H. E., and Yantis, S. (2004). Visual attention: bottom-up versus top-down. *Curr. Biol.* 14, R850–R852. doi: 10.1016/j.cub.2004.09.041

de Fockert, J. W., Rees, G., Frith, C. D., and Lavie, N. (2001). The role of working memory in visual selective attention. *Science* 291, 1803–1806,

De la Torre, G. G., Barroso, J. M., León-Carrión, J., Mestre, J. M., and Bozal, R. G. (2015). Reaction time and attention: toward a new standard in the assessment of ADHD? A pilot study. *J. Atten. Disord.* 19, 1074–1082. doi: 10.1177/1087054712466440

Della Sala, S., Laiacona, M., Spinnler, H., and Ubezio, C. (1992). A cancellation test: its reliability in assessing attentional deficits in Alzheimer's disease. *Psychol. Med.* 22, 885–901. doi: 10.1017/S0033291700038460

Demeyere, N., Riddoch, M. J., Slavkova, E. D., Bickerton, W. L., and Humphreys, G. W. (2015). The Oxford cognitive screen (OCS): validation of a stroke-specific short cognitive screening tool. *Psychol. Assess.* 27, 883–894. doi: 10.1037/pas0000082

Eglin, M., Robertson, L. C., and Knight, R. T. (1989). Visual search performance in the neglect syndrome. *J. Cogn. Neurosci.* 1, 372–385. doi: 10.1162/jocn.1989.1.4.372

Facchin, A., Simioni, M., Maffioletti, S., and Daini, R. (2023). Broken ring enVision search (BReViS): a new clinical test of attention to assess the effect of layout and crowding on visual search. *Brain Sci.* 13:494. doi: 10.3390/brainsci13030494

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G*power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Hatta, T., Masui, T., Ito, Y., Ito, E., Hasegawa, Y., and Matsuyama, Y. (2004). Relation between the prefrontal cortex and cerebro-cerebellar functions: evidence from the results of stabilometrical indexes. *Appl. Neuropsychol.* 11, 153–160. doi: 10.1207/s15324826an1103_3

Hultsch, D., Hertzog, C., Small, B. J., and Dixon, R. A. (1999). Use it or lose it? Engage lifestyle as a buffer of cognitive decline in aging? *Psychol. Aging* 14, 245–263. doi: 10.1037/0882-7974.14.2.245

Jones, R. N., Yang, F. M., Zhang, Y., Kiely, D. K., Marcantonio, E. R., and Inouye, S. K. (2006). Does educational attainment contribute to risk for delirium? A potential role for cognitive reserve. *Journal of Gerontology: Medical Sciences.* 61, 1307–1311. doi: 10.1093/gerona/61.12.1307

Ku, Y. (2018). Selective attention on representations in working memory: cognitive and neural mechanisms. *PeerJ* 6:e4585. doi: 10.7717/peerj.4585

Lee, S., Kawachi, I., Berkman, L. F., and Grodstein, F. (2003). Education, other socioeconomic indicators, and cognitive function. *Am. J. Epidemiol.* 157, 712–720. doi: 10.1093/aje/kwg042

Lindenberger, U., and Baltes, P. B. (1997). Intellectual functioning in old and very old age: cross-sectional results from the Berlin aging study. *Psychol. Aging* 12, 410–432. doi: 10.1037/0882-7974.12.3.410

Luck, S. J., Hillyard, S. A., Mangun, G. R., and Gazzaniga, M. S. (1989). Independent hemispheric attentional systems mediate visual search in split-brain patients. *Nature* 342, 543–545. doi: 10.1038/342543a0

Lyketsos, C. G., Chen, L., and Anthony, J. C. (1999). Cognitive decline in adulthood: an 11.5 year follow-up of the Baltimore epidemiological catchment area study. *Am. J. Psychiatry* 156, 58–65. doi: 10.1176/ajp.156.1.58

Madden, D. J., Spaniol, J., Whiting, W. L., Bucur, B., Provenzale, J. M., Cabeza, R., et al. (2007). Adult age differences in the functional neuroanatomy of visual attention: a combined fMRI and DTI study. *Neurobiol. Aging* 28, 459–476. doi: 10.1016/j.neurobiolaging.2006.01.005

Malavita, M. S., Vidyasagar, T. R., and McKendrick, A. M. (2017). The effect of aging and attention on visual crowding and surround suppression of perceived contrast threshold. *Invest. Ophthalmol. Vis. Sci.* 58, 860–867. doi: 10.1167/iovs.16-20632

Marioni, R. E., van den Hout, A., Valenzuela, M. J., Brayne, C., Matthews, F. E., Function, M. R. C. C., et al. (2012). Active cognitive lifestyle associates with cognitive recovery and a reduced risk of cognitive decline. *J. Alzheimers Dis.* 28, 223–230. doi: 10.3233/JAD-2011-110377

Mesulam, M.-M. (2000). *Principles of behavioral and cognitive neurology*. Oxford, UK: Oxford University Press.

Moore, T., and Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annu. Rev. Psychol.* 68, 47–72. doi: 10.1146/annurev-psych-122414-033400

Mueller, V., Brehmer, Y., von Oertzen, T., Li, S. C., and Lindenberger, U. (2008). Electrophysiological correlates of selective attention: a lifespan comparison. *BMC Neurosci.* 9:18. doi: 10.1186/1471-2202-9-18

Oberauer, K. (2019). Working memory and attention - a conceptual analysis and review. *J. Cogn.* 2:36. doi: 10.5334/joc.58

Park, M. O., and Lee, S. H. (2019). Effect of a dual-task program with different cognitive tasks applied to stroke patients: a pilot randomized controlled trial. *Neuro Rehabil.* 44, 239–249. doi: 10.3233/NRE-182563

Pegoraro, S., Facchin, A., Luchesa, F., Rolandi, E., Guaita, A., Arduino, L. S., et al. (2024). The complexity of Reading revealed by a study with healthy older adults. *Brain Sci.* 14:230. doi: 10.3390/brainsci14030230

Petersen, S. E., and Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* 35, 73–89. doi: 10.1146/annurev-neuro-062111-150525

Plebanek, D. J., and Sloutsky, V. M. (2019). Selective attention, filtering, and the development of working memory. *Dev. Sci.* 22:e12727. doi: 10.1111/desc.12727

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Posner, M. I., and Petersen, S. E. (1990). The attention system of the human brain. *Annu. Rev. Neurosci.* 13, 25–42. doi: 10.1146/annurev.ne.13.030190.000325

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372

Reigal, R. E., Barrero, S., Martín, I., Morales-Sánchez, V., Juárez-Ruiz de Mier, R., and Hernández-Mendo, A. (2019). Relationships between reaction time, selective attention, physical activity, and physical fitness in children. *Front. Psychol.* 10:2278. doi: 10.3389/fpsyg.2019.02278

Rubia, K., Hyde, Z., Halari, R., Giampietro, V., and Smith, A. (2010). Effects of age and sex on developmental neural networks of visual-spatial attention allocation. *NeuroImage* 51, 817–827. doi: 10.1016/j.neuroimage.2010.02.058

Senger, C., Margarido, M. R. R. A., De Moraes, C. G., De Fendi, L. I., Messias, A., and Paula, J. S. (2017). Visual search performance in patients with vision impairment: a systematic review. *Curr. Eye Res.* 42, 1561–1571. doi: 10.1080/02713683.2017.1338348

Shamsi, F., Liu, R., and Kwon, M. (2022). Foveal crowding appears to be robust to normal aging and glaucoma unlike parafoveal and peripheral crowding. *J. Vis.* 22:10. doi: 10.1167/jov.22.8.10

Stern, Y., Haback, C., Moeller, J., Scarmeas, N., Anderson, K. E., Hilton, H. J., et al. (2005). Brain networks associated with cognitive reserve in healthy young and old adults. *Cerb. Cort.* 15, 394–402. doi: 10.1093/cercor/bhh142

Strober, L., DeLuca, J., Benedict, R. H., Jacobs, A., Cohen, J. A., Chiaravalloti, N., et al. (2019). Symbol digit modalities test: a valid clinical trial endpoint for measuring cognition in multiple sclerosis. *Mult. Scler.* 25, 1781–1790. doi: 10.1177/1352458518808204

Tavakol, M., and Dennick, R. (2011). Making sense of Cronbach's alpha. *Int. J. Med. Educ.* 2, 53–55. doi: 10.5116/ijme.4dfb.8dfd

Tun, P. A., and Lachman, M. E. (2008). Age differences in reaction time and attention in a national telephone sample of adults: education, sex, and task complexity matter. *Dev. Psychol.* 44, 1421–1429. doi: 10.1037/a0012845

Turatto, M., Benso, F., Facoetti, A., Galfano, G., Mascetti, G. G., and Umiltà, C. (2000). Automatic and voluntary focusing of attention. *Percept. Psychophys.* 62, 935–952. doi: 10.3758/BF03212079

Uttl, B., and Pilkenton-Taylor, C. (2001). Letter cancellation performance across the adult life span. *Clin. Neuropsychol.* 15, 521–530. doi: 10.1076/clin.15.4.521.1881

Utz, K. S., Hankeln, T. M., Jung, L., Lammer, A., Waschbisch, A., Lee, D. H., et al. (2013). Visual search as a tool for a quick and reliable assessment of cognitive functions in patients with multiple sclerosis. *PLoS One* 8:e81531. doi: 10.1371/journal.pone.0081531

van Es, D. M., Theeuwes, J., and Knapen, T. (2018). Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. *eLife*. 7:e36928. doi: 10.7554/eLife.36928

Verghese, P. (2001). Visual search and attention: a signal detection theory approach. *Neuron* 31, 523–535. doi: 10.1016/S0896-6273(01)00392-0

Veríssimo, J., Verhaeghen, P., Goldman, N., Weinstein, M., and Ullman, M. T. (2022). Evidence that ageing yields improvements as well as declines across attention and executive functions. *Nat. Hum. Behav.* 6, 97–110. doi: 10.1038/s41562-021-01169-7

Weintraub, S., and Mesulam, M. M. (1988). Visual Hemispatial inattention: stimulus parameters and exploratory strategies. *J. Neurol. Neurosurg. Psychiatry* 51, 1481–1488. doi: 10.1136/jnnp.51.12.1481

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends Cogn. Sci.* 7, 70–76. doi: 10.1016/S1364-6613(02)00024-4

Wolfe, J. M. (2020). Visual search: how do we find what we are looking for? *Annu. Rev. Vis. Sci.* 6, 539–562. doi: 10.1146/annurev-vision-091718-015048

Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J., and Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vis. Res.* 42, 2985–3004. doi: 10.1016/S0042-6989(02)00388-7

Yehezkel, O., Sterkin, A., Lev, M., and Polat, U. (2015). Crowding is proportional to visual acuity in young and aging eyes. *J. Vis.* 15:23. doi: 10.1167/15.8.23

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 July 2024
DOI 10.3389/fpsyg.2024.1412560

OPEN ACCESS

EDITED BY
Alessio Facchin,
Magna Graecia University, Italy

REVIEWED BY
Elena Cavallini,
University of Pavia, Italy
Antonia Meyer,
University Hospital of Basel, Switzerland

*CORRESPONDENCE
Federica Rossetto
✉ frossetto@dongnocchi.it

RECEIVED 05 April 2024
ACCEPTED 18 July 2024
PUBLISHED 30 July 2024

CITATION
Isernia S, Cacciatore DM, Rossetto F, Ricci C and Baglio F (2024) Reliability and minimal detectable change of the Yoni task for the theory of mind assessment. *Front. Psychol.* 15:1412560. doi: 10.3389/fpsyg.2024.1412560

Reliability and minimal detectable change of the Yoni task for the theory of mind assessment

Sara Isernia[1], Diego Michael Cacciatore[1], Federica Rossetto[1]*, Cristian Ricci[2] and Francesca Baglio[1]

[1]IRCCS Fondazione Don Carlo Gnocchi ONLUS, Milan, Italy, [2]Africa Unit for Transdisciplinary Health Research (AUTHeR), North-West University, Potchefstroom, South Africa

Introduction: The Theory of Mind (ToM) assessment is becoming essential to evaluate the response to a social cognition intervention and to monitor the progression of social abilities impairment in atypical conditions. In the Italian setting, the Yoni task has been recently validated in its short version (the Yoni-48 task) to evaluate ToM in the clinical setting. The present study aimed to verify the test-retest reliability and the Minimal Detectable Change (MDC) of the Yoni-48 task.

Methods: The Yoni-48 task was administered to 229 healthy adults at two evaluation sessions 3 weeks apart (mean days between sessions = 20.35 ± 1.75) by a psychologist. The test-retest reliability of the Yoni-48 task accuracy and response time was tested by the Intraclass Correlation Coefficient ($ICC_{2,1}$, two-way random model, absolute agreement type). Then, the MDC_{95} and MDC_{90} were computed based on the standard error of measurement. Finally, the 95% limits of agreement were plotted (LOA plot) to visualize the difference and mean score of each pair of measurements.

Results: The total Yoni-48 task accuracy, but not the response time score, showed a high ICC (>0.80), with an MDC of 0.10. By plotting the LOA plot for the accuracy score no systematic trends were observed.

Discussion: This evidence will support the adoption of the Yoni task in longitudinal designs.

KEYWORDS

social cognition, mentalizing, test-retest, reliability, rehabilitation

1 Introduction

Social cognition is a complex set of abilities enabling the detection and processing of social stimuli from the environment. It allows adequate social behavioral response (Frith, 2008) and successful social relationships, which are essential for physical and psychological well-being (Umberson and Montez, 2010). A core component of social cognition is the Theory of Mind (ToM) or mentalizing, the capacity to infer own and others' mental states (i.e., emotions, beliefs, and intentions) to predict behavior (Premack and Woodruff, 1978; Wimmer and Perner, 1983). ToM has a multidimensional and multilevel nature. Especially, it consists of an affective (*hot*) and cognitive (*cold*) component, which involves the understanding of affective (emotions) and cognitive (beliefs, intentions, thoughts) mental states, respectively (Brothers and Ring, 1992; Abu-Akel and Shamay-Tsoory, 2011). Also, two different levels of complexity of ToM reasoning have been highlighted (Shamay-Tsoory et al., 2005; Kalbe et al., 2010)

referring to the first-order ToM, the capacity to represent another person's emotions/beliefs/intentions, and the second-order ToM, the ability to attribute one person's belief about another person's mental state (Happè, 2021).

In recent years, the assessment of ToM in the clinical setting is become essential. ToM deficits are frequently considered markers of social maladaptation linked to a broad range of developmental, psychiatric, and neurological disorders (Bora and Pantelis, 2013; Plana et al., 2014; Cotter et al., 2016). Moreover, ToM performance may serve as a marker of neural deterioration and disease progression. There is evidence that social cognitive impairment characterizes the early stage of many clinical conditions, including the early stage of dementia (Bora et al., 2015; Rossetto et al., 2020, 2022; Yi et al., 2020), and that ToM deficits get worse with the progression of the disease (Bora et al., 2015, 2016), leading to poor social and occupational functioning and reduced quality of life. For this reason, ToM measures have to be included in the neuropsychological battery to monitor the progression of neurocognitive symptoms (as suggested by the Diagnostic and Statistical Manual of Mental Disorders, 5th edition, American Psychiatric Association, 2013) and to customize the rehabilitation treatment strategies. In fact, social cognition rehabilitation activities may be integrated into cognitive interventions for several neurological and neurodegenerative conditions (Henry et al., 2016), given the flourishing evidence on social abilities impairment in these populations. Finally, ToM measures may be adopted to assess the response to a social cognition intervention. Specific rehabilitation programs targeted to enhance social cognition abilities have been implemented and proposed for people with neuropsychiatric and neurological diseases, such as schizophrenia (d'Arma et al., 2021), traumatic brain injury (Togher et al., 2023), and Multiple Sclerosis (d'Arma et al., 2023). However, few ToM measures have been tested for longitudinal evaluations and to be adopted in rehabilitation settings. In fact, in this context, some psychometric properties, such as the test-retest reproducibility evidence and the estimation of score responsiveness, such as the minimal detectable change, are needed for a good interpretation of the rehabilitation trajectories and responses. Finally, changes in ToM competencies are frequently assessed longitudinally through long and time-consuming composite batteries that attempt to understand the complex nature of the construct.

The Yoni task (Shamay-Tsoory et al., 2007) has been recently validated and standardized in its 48-item short version (the Yoni-48 task) (Isernia et al., 2022a,b) for widespread use in clinical settings, also for a longitudinal approach. The advantage of this test is the multidimensional and multi-level assessment of ToM by evaluating separately cognitive and affective domains, and first- and second-order mental states attribution. Moreover, it was conceived as a digital measure (Koo and Vizer, 2019): it is administered in a computerized way, allowing the simultaneous collection of both accuracy and response time scores. Especially, each item is scored based on a corrected/uncorrected answer and on the seconds taken to answer. Importantly, the Yoni task consists of visual stimuli minimizing the influence of language, memory, and executive function on the subject's performance. Moreover, the adoption of the Yoni task in the assessment of ToM in the clinical population has been supported by previous studies. Especially, it has been demonstrated to effectively detect ToM difficulties in localized brain lesions conditions (Abu-Akel and Shamay-Tsoory, 2011), schizophrenia, Parkinson's Disease, and Mild Cognitive Impairment (Rossetto et al., 2018). Based on this previous evidence, the Yoni task is suggested to be suitable for the clinical setting, such as for supporting neuropsychological assessment.

However, a study testing the reproducibility over time of the Yoni task is needed to provide further proof of reliability. Importantly, an estimate of the Minimal Detectable Change (MDC), that is the minimal magnitude of change beyond which the change is real rather than a random measurement error, is needed for the adoption of the tool in the longitudinal contexts. The MCD is commonly computed for measures of motor functions, which are widely adopted for the monitoring of the performance after a rehabilitation program (e.g., Watson and Petrie, 2010; Lee et al., 2013; Palmer et al., 2017; Negrete et al., 2021). However, it is rarely estimated for cognitive measures (Blackwood et al., 2021; Webb et al., 2022; Chiu et al., 2023), and, to our knowledge, it has been never computed for social cognition tools.

The present study aimed to verify the reproducibility of the Yoni-48 task by estimating the test-retest reliability and the MDC value.

2 Materials and methods

This is a prospective study conducted from November 2022 to December 2023 at the IRCCS Don Gnocchi Foundation (Milan, Italy). The research has been reviewed and approved by the Don Gnocchi Foundation Ethics Committee.

2.1 Participants

Participants were recruited from the university courses (students of Professional Education; Psychology; Nurse; Psychomotricity) and the staff (technical staff; health professionals; interns) of the IRCCS Don Gnocchi Foundation, Santa Maria Nascente Center of Milan (Italy). Inclusion criteria considered to enroll participants was age > 18. Also, the following exclusion criteria were considered as well: (i) presence of neurological and/or psychiatric conditions; (ii) presence of visual and hearing disability able to affect the performance of the task; (iii) presence of pharmacological therapy affecting the evaluation session.

Participation in the study was voluntary and subjects did not receive pecuniary compensation for their involvement in the research.

2.2 Measures

2.2.1 The Yoni-48 task

The Yoni task is a computerized measure of ToM originally developed by Shamay-Tsoory and Aharon-Peretz (2007). The task is composed of visual static stimuli, in which a face ("Yoni") appears at the center of the screen, surrounded by 4 elements (fruits, characters, animals…). For each stimulus, based on a written instruction on the top of the screen, the subject is invited to click on the element Yoni refers to, having not more than 60 s maximum per item. Therefore, the subjects are required to infer cognitive (cognitive ToM items: e.g., "*Yoni is thinking of…*") and affective (affective ToM items: e.g., "*Yoni loves…*") mental states of Yoni. The gaze direction and the facial expression of Yoni are informative cues to choose the right answer. Also, control stimuli are included in which the subject is invited to

perform a physical inference (control items: e.g., "*Yoni is close to*…"). Moreover, stimuli show two levels of ToM recursive thinking, assessing first- (e.g., "*Yoni is thinking of* …") and second-order ToM (e.g., "*Yoni is thinking about the fruit that … wants*"), respectively. In this study, the Italian version of the task (48-item; Isernia et al., 2022a,b) was administered. This version is constituted of 42 ToM and 6 control items. The ToM items are divided into 21 affective and 21 cognitive ToM; 16 first- and 26 second-order ToM items. The accuracy and response time scores have been separately computed based on Italian scoring instructions and adjusted for demographic variables, such as sex, age, and education (Isernia et al., 2022b). The following composite scores have been calculated: accuracy composite score (ACC, range 0–1), and response time composite score (RT, range 0–1).

2.3 Procedure

The Yoni task was administered at two evaluation sessions (test and retest sessions) three weeks apart (mean days between sessions = 20.35 ± 1.75) by a psychologist. The evaluation sessions were conducted in the same setting using the same technological device to perform the task (Figure 1). Within the test session, participant demographics were also collected.

2.4 Statistical analysis

Statistical analysis was performed using IBM SPSS software (version 28.0) and R (version 4.1.2). Descriptive statistics (frequencies, means, medians, standard errors, and standard deviations) were reported to detail the demographics of the participants group and their performance in the Yoni task at the test and retest sessions.

Before reliability analyses, outliers were identified considering the Yoni task performance under 2 standard deviations from the norm (Isernia et al., 2022b) (see Figure 2). Then, to observe the extent of the Yoni task score fluctuation (practice effect) between the test and re-test session, the effect size (point-biserial correlation coefficient, r_{pbs}) of paired-sample comparison (Wilcoxon rank test) between test and retest performance was extracted. Also, the correlation between test-retest Δ change and the mean of the two assessments ($M_{assessment}$) was run, and the 95% limits of agreement were plotted (LOA plot, Bland and Altman, 1986) to visualize the difference and mean score of each pair of measurements.

The repeatability of the Yoni task accuracy and response time was tested by the intraclass correlation coefficient ($ICC_{2,1}$, two-way random model, absolute agreement type). An ICC score ≥ 0.80, 0.79–0.60, and <0.59 was interpreted as a high, moderate, and poor agreement, respectively.

Then, the minimal detectable change (MDC) value of the Yoni task scores was computed as agreement parameters to be used to determine consistent improvement or decrement in the ToM ability, net to the measure oscillations. To this purpose, the standard error of measurement (SEM), the MDC_{95}, and MDC_{90} were calculated using the following formula:

$$SEM = SD\ all\ testing\ score\ x\ \sqrt{1 - ICC}$$

$$MDC_{95} = 1.96\ x\ \sqrt{2}\ x\ SEM$$

$$MDC_{90} = 1.65\ x\ \sqrt{2}\ x\ SEM$$

Then, the amount of random measurement error (MCD_{95}%; MCD_{90}%) was computed by dividing MCD_{95}/MCD_{90} by the maximum score and multiplying it by 100.

As additional analyses, to confirm the validity and inter-item reliability of the Yoni task, internal consistency (Cronbach α), split-half reliability, Pearson reliability, mean infit and outfit were computed at T1 and T2. Moreover, construct validity was assessed at T1 using a confirmatory factor analysis. Firstly, the factorial scores representing the construct have been computed separately considering the affective and cognitive items. Afterward, the Spearman correlations were reported to portray the association between the item and the factor score. The cfa function of the R

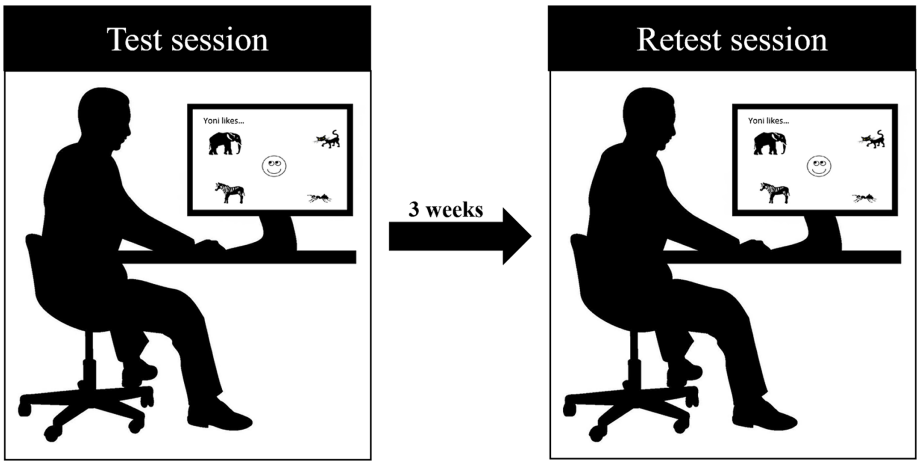


FIGURE 1
The study procedure.

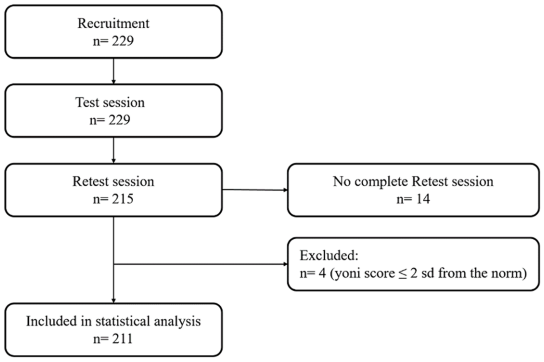


FIGURE 2
Flow chart of the study.

TABLE 1 Demographics of the participants in the study.

	Participants in test and retest sessions	Participants included in the analysis
N	215	211
Sex (Ma:F)	55:162	53:158
Age (*M* ± sd)	25.48 ± 9.18	25.53 ± 9.24
Education	13.99 ± 2.10	13.99 ± 2.11
Occupation		
Students (%)	73	71
Workers (%)	27	29

F, females; M, mean; Ma, males; sd, standard deviation.

lavaan package was used to perform the confirmatory factor analysis and, the option ordered = TRUE was used to consider the categorical nature of the items (Rosseel, 2012).

3 Results

3.1 Participants

A total of 229 healthy adults took part in the research. Among these, 215 participants attended both the test and retest sessions (Table 1). Four people were identified as outliers and were excluded from the analysis since they reported a Yoni task performance far from the norm (z score ≤ 2 sd of the normative population; Isernia et al., 2022b). In total, 211 participants were included in the analyses [53 males, mean age = 25.53 ± 9.24; mean education (y) = 13.99 ± 2.11]. Figure 2 depicts the flow chart of the study.

3.2 The Yoni task performance in the test and retest sessions

Tables 2 and 3 show the performance of participants at the Yoni task in the test and retest sessions. Both the accuracy and the response time scores were high in the test session and tended to increase in the retest session (see Figure 3). The Wilcoxon W test reported a statistically significant difference between the two sessions' performance in all scores except for the first-order and cognitive accuracy scores. The effect size (r_{pbs}) suggested a slight practice effect in the accuracy performance and a moderate effect in the response time.

By plotting the test-retest Δ change against the mean score of the assessments (Bland–Altman plot, Figure 4) for ACC and RT scores, no systematic trends were observed.

3.3 The Yoni task reliability

3.3.1 Repeatability results

Table 4 reports the ICC values of the Yoni task scores. Results suggested a good reliability of all the accuracy scores except for the ToM first-order score. Specifically, the ToM total (ACC), second-order, and cognitive accuracy scores showed a high repeatability (ICC > 0.80), while the ToM affective score revealed a moderate test-retest reliability. The ToM first-order score, instead, showed poor repeatability (ICC < 0.59). Concerning the response time scores, we observed poor reliability (ICC < 0.59) both in the total response time score (RT) and sub-scores (Table 5).

3.3.2 Agreement parameters

The MDC values of accuracy and response time scores suggested an acceptable-to-excellent random measurement error (Tables 4, 5). Especially, the accuracy total score, which reported also high repeatability, showed an MDC% equal to 11.48 for a 95% confidence level, and equal to 9.66 for a 90% confidence level. In particular, a fluctuation >/< 0.10 in the ACC score can be interpreted as a consistent improvement/decrement in the ToM performance.

3.3.3 Inter-item reliability, item discrimination ability and construct validity

To further explore reliability, the Yoni task internal consistency, split-half reliability, and item discrimination ability were explored at T1 and T2. The Yoni task showed a high inter-item reliability at both times: an internal consistency Cronbach's α = 0.80 at T1 and T2, and a good median split-half reliability in both times (T1: ϱ_{SP} = 0.81 and a 95% HDI = 0.73–0.86; T2: ϱ_{SP} = 0.81 and a 95% HDI = 0.72–0.86). Also, the dichotomous Rash model analysis revealed a Pearson reliability of the test equal to 0.613 at T1 and 0.55 at T2. Finally, the items showed a mean infit equal to 0.93 ± 0.27 at T1 and 0.93 ± 0.28 at T2, and a mean outfit equal to 0.99 ± 0.75 at T1 and 1.06 ± 0.89 at T2. Construct validity was confirmed by the confirmatory factor analysis: ϱ Spearman's correlation coefficient reported significant associations between the affective ToM and cognitive ToM latent factors and affective and cognitive items, respectively (see Supplementary Table S1). Three items reported weak/absent associations with the latent factor: items 8, 13, and 38.

4 Discussion

The Yoni-48 task has been proposed as a digital tool for the assessment of ToM, which has been recently validated for the Italian population (Isernia et al., 2022b) and may be suitable to be adopted as an outcome measure in social cognition interventions. Its digital administration complies with the recent advantage of digital neuropsychology (Bilder, 2011), allowing the norm-based

TABLE 2 Comparison between the Yoni task accuracy in test and retest sessions.

	T1			T2			Δchange			$M_{assessments}$			Wilcoxon rank		
	M	SE	SD	*M*	SE	SD	*M*	SE	SD	*M*	SE	SD	*W*	*p*	r_{pbs}
ACC	0.88	0.01	0.10	0.89	0.01	0.09	0.01	0.00	0.08	0.88	0.00	0.09	7747.00	0.004	0.268
1ORD	15.74	0.05	0.77	15.85	0.04	0.54	0.11	0.05	0.75	15.79	0.04	0.56	367.50	0.053	0.392
2ORD	21.10	0.27	4.00	21.55	0.27	3.98	0.45	0.21	3.06	21.29	0.25	3.68	7278.50	0.007	0.249
AFF	18.31	0.15	2.25	18.72	0.15	2.22	0.41	0.15	2.17	18.50	0.13	1.96	5758.00	0.003	0.290
COG	18.56	0.18	2.70	18.73	0.17	2.45	0.18	0.14	1.99	18.63	0.16	2.38	4737.00	0.146	0.150

ACC, ToM accuracy composite score; AFF, affective ToM score; COG, cognitive ToM score; M, mean; $r_{pbs,}$ r point-biserial correlation coefficient; SE, standard error; SD, standard deviation; W, Wilcoxon test; 1ORD, first-order ToM score; 2ORD, second-order ToM score.

TABLE 3 Comparison between the Yoni task response time in test and retest sessions.

	T1			T2			Δchange			$M_{assessments}$			Wilcoxon rank		
	M	SE	SD	*M*	SE	SD	*M*	SE	SD	*M*	SE	SD	*W*	*p*	r_{pbs}
RT	0.89	0.00	0.04	0.91	0.00	0.04	0.03	0.00	0.04	0.90	0.00	0.03	19079.00	<0.001	0.706
1ORD	5.69	0.17	2.44	4.38	0.12	1.81	−1.30	0.17	2.47	5.04	0.12	1.76	4494.00	<0.001	0.598
2ORD	11.60	0.23	3.41	9.39	0.20	2.96	−2.20	0.23	3.42	10.58	0.18	2.70	3711.00	<0.001	0.668
AFF	9.95	0.22	3.15	7.54	0.17	2.44	−2.41	0.22	3.22	8.81	0.16	2.33	2916.00	<0.001	0.739
COG	9.04	0.19	2.80	7.67	0.18	2.61	−1.36	0.19	2.80	8.41	0.16	2.31	5321.00	<0.001	0.524

AFF, affective ToM score; COG, cognitive ToM score; M, mean; $r_{pbs,}$ r point-biserial correlation coefficient; RT, response time composite score; SE, standard error; SD, standard deviation; W, Wilcoxon test; 1ORD, first-order ToM score; 2ORD, second-order ToM score.

administration of test batteries via computers, tablets and mobiles (Koo and Vizer, 2019). Especially, the Yoni task has been conceived as a computerized tool, able to facilitate agile data recording and scoring.

The present study tested the reproducibility of the Yoni-48 task to evaluate its reliability for the assessment of social cognition in longitudinal contexts, such as in the pre- and post-evaluation of rehabilitation and intervention programs.

First, the accuracy score (ACC) of the Yoni task showed good test-retest reliability, demonstrating high stability over time and minimal learning effects. To date, only a few studies investigated the test-retest reliability of ToM measures, reporting mixed results. In this regard, the Yoni-48 task revealed a higher reproducibility than other ToM tools. Especially, it showed slightly higher reliability than ToM measures already estimated as highly reproducible (Yeh et al., 2021), such as the Hinting Task (Corcoran et al., 1995) and the Faux-pas test (Stone et al., 1998, 2003). Also, the Yoni-48 task reliability was far greater than other widely used social cognition tools, such as the False Beliefs test and the Story tests (Chen et al., 2017). Based on the study of Altschuler and Faja (2022), the Yoni-48 task reliability outperformed also the second-order false belief test (Muris et al., 1999) and the Social Attribution Task (Klin, 2000), which demonstrated good reproducibility in a cohort of 7 to 11 year-old children within autism spectrum disorder. Finally, the stability over time of the Yoni-48 task was equally good as the Reading the Mind in the Eyes Test, as reported by Vellante et al. (2013), which is one of the most used ToM tests in the Italian context.

By considering the Yoni-48 sub-scores, we observed different levels of test-retest reliability. Especially, the second-order ToM score showed a higher stability than the first-order score. This result might be related to the greater sensitivity of the second-order items than the first-order ones, as suggested by previous works (Isernia et al., 2022a,b). Also, although both cognitive and affective ToM scores were fairly stable, the cognitive ToM score showed a higher reliability. This result may be likely explained by the major relevance of visual cues in the affective than cognitive ToM items, which required subjects to capture the affective mental states based on the facial expressions and could be more influenced by visual processing and related habituation effects (Breiter et al., 1996; Pirastru et al., 2023).

Although the accuracy score of the Yoni-48 task has been found to be reliable, the response time score (RT) did not reach acceptable stability. In fact, our findings suggested that the RT score was affected by the learning effect and increased over time. This result was expected and may be related to the familiarity with the stimuli modality and the task instructions, which influenced the subjects' processing speed (Balas et al., 2007). Globally, this evidence is suggestive of the reliability of the Yoni-48 task and its application as a reliable ToM measure in longitudinal design studies by considering the accuracy and not the response time score. Especially, the global accuracy score (ACC) would be used in future studies to monitor ToM ability. Although we found a high reproducibility of the second-order and cognitive ToM accuracy score, focusing on only one sub-score (such as cognitive ToM and not affective ToM) may be avoided unless under a strict theory-driven hypothesis.

After exploring the reliability of the test, the minimal detectable change was estimated to obtain a measure of the minimal magnitude change of the tool. This value will be useful to capture significant variations in the ToM performance that may not be associated with the measurement error (de Vet et al., 2006). Especially, our findings indicated that an oscillation of 0.11 points in the ACC score should be interpreted as an informative change and may suggest a significant increment/decrement of the performance over time. This datum will be considered as a reference point for the ToM monitoring, training, and rehabilitation.

This study is not without limitations. Our participants were healthy young adults with a high ToM ability. Future studies may include people with ToM difficulties to give clues about the reliability of the Yoni-48 task in clinical populations (e.g., schizophrenia, autism

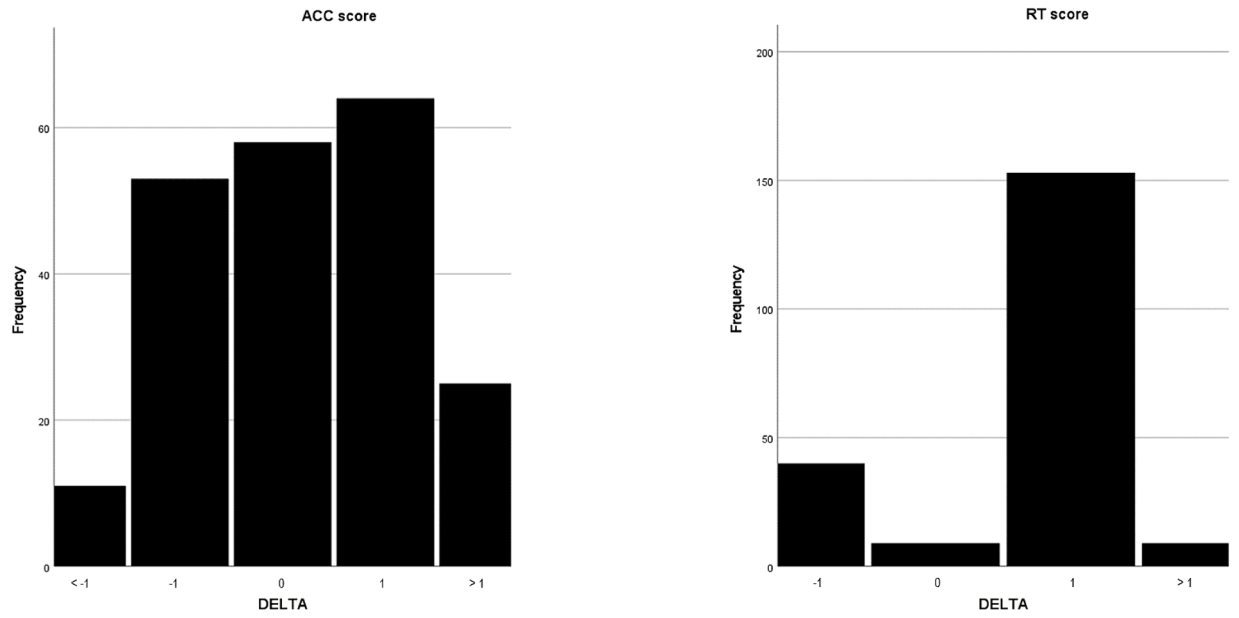


FIGURE 3
Delta change trend of ACC and RT scores.

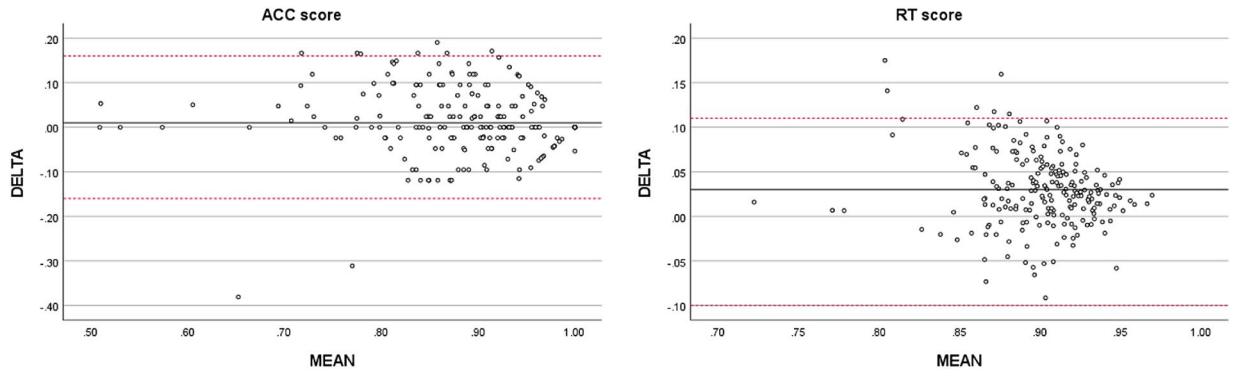


FIGURE 4
Bland-Altman plot of differences in scores against the mean scores of the two assessments.

TABLE 4 Reliability and minimal detectable change of the Yoni-48 task accuracy scores.

	ICC	SEM	MDC90	MDC95	MDC90%	MDC95%
ACC	0.81	0.04	0.10	0.11	9.66	11.48
1ORD	0.53	0.45	1.05	1.24	6.55	7.78
2ORD	0.82	1.69	3.95	4.69	15.19	18.05
AFF	0.69	1.24	2.90	4.45	13.83	16.42
COG	0.82	1.09	2.55	3.03	12.14	14.42

AFF, affective ToM score; COG, cognitive ToM score; ICC, Intraclass correlation coefficient; SEM, Standard error mean, MDC, minimal detectable change; 1ORD, first-order ToM score; 2ORD, second-order ToM score.

spectrum disorder, neurological conditions). Also, further description of the demographic characteristics of the participants, such as the specific work activity, marital status, and ethnicity, should have been collected, as well as subclinical conditions such as depression and autism spectrum symptoms to test the impact of these variables on the Yoni task performance. Moreover, our participants' group was composed of a higher rate of females than males, and gender differences were not considered. Finally, our results on minimal detectable change (ACC score change of 0.11) may be interpreted solely as a reference point to capture Yoni-48 real changes and not

TABLE 5 Reliability and minimal detectable change of the Yoni-48 task RT scores.

	ICC	SEM	MDC90	MDC95	MDC90%	MDC95%
RT	0.52	0.03	0.06	0.08	6.47	7.68
1ORD	0.44	1.59	3.71	4.41	6.18	7.35
2ORD	0.51	2.23	5.20	6.18	8.67	10.30
AFF	0.40	2.16	5.05	6.00	8.42	10.00
COG	0.58	1.75	4.09	4.86	6.82	8.10

AFF, affective ToM score; COG, cognitive ToM score; ICC, Intraclass correlation coefficient; SEM, Standard error mean, MDC, minimal detectable change; 1ORD, first-order ToM score; 2ORD, second-order ToM score.

clinically meaningful changes. For this latter purpose, future research may include a measure of the health status and estimate the minimal clinically important difference in a clinical population target.

5 Conclusion

In conclusion, to our knowledge, this is the first study that estimated the test-retest reliability of the Yoni task and computed the minimal detectable change for a ToM measure. This evidence will support future studies on social cognition trainings and will sustain the interpretation of the Yoni task scores in longitudinal designs.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Don Gnocchi Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

SI: Conceptualization, Formal analysis, Methodology, Writing – original draft. DC: Data curation, Writing – original draft. FR: Data curation, Writing – review & editing. CR: Methodology, Writing – review & editing. FB: Conceptualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Italian Ministry of Health (Ricerca Corrente 2022-2024).

Acknowledgments

The authors thanks all the participants who took part in the research.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1412560/full#supplementary-material

References

Abu-Akel, A., and Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* 49, 2971–2984. doi: 10.1016/j.neuropsychologia.2011.07.012

Altschuler, M. R., and Faja, S. (2022). Brief report: test-retest reliability of cognitive, affective, and spontaneous theory of mind tasks among school-aged children with autism Spectrum disorder. *J. Autism Dev. Disord.* 52, 1890–1895. doi: 10.1007/s10803-021-05040-6

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM V)*. Washington, DC, USA: American Psychiatric.

Balas, B., Cox, D., and Conwell, E. (2007). The effect of real-world personal familiarity on the speed of face information processing. *PLoS One* 2:e1223. doi: 10.1371/journal.pone.0001223

Bilder, R. M. (2011). Neuropsychology 3.0: evidence-based science and practice. *J. Int. Neuropsychol. Soc.* 17, 7–13. doi: 10.1017/s1355617710001396

Blackwood, J., Rybicki, K., and Huang, M. (2021). Cognitive measures in older cancer survivors: an examination of validity, reliability, and minimal detectable change. *J. Geriatr. Oncol.* 12, 146–151. doi: 10.1016/j.jgo.2020.06.015

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310

Bora, E., and Pantelis, C. (2013). Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. *Schizophr. Res.* 144, 31–36. https://10.1016/j.schres.2012.12.013. doi: 10.1016/j.schres.2012.12.013

Bora, E., Velakoulis, D., and Walterfang, M. (2016). Meta-analysis of facial emotion recognition in behavioral variant frontotemporal dementia: comparison with Alzheimer disease and healthy controls. *J. Geriatr. Psychiatry Neurol.* 29, 205–211. doi: 10.1177/0891988716640375

Bora, E., Walterfang, M., and Velakoulis, D. (2015). Theory of mind in behavioural-variant frontotemporal dementia and Alzheimer's disease: a meta-analysis. *J. Neurol. Neurosurg. Psychiatry* 86, 714–719. doi: 10.1136/jnnp-2014-309445

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17, 875–887. doi: 10.1016/s0896-6273(00)80219-6

Brothers, L., and Ring, B. (1992). A neuroethological framework for the representation of minds. *J. Cogn. Neurosci.* 4, 107–118. doi: 10.1162/jocn.1992.4.2.107

Chen, K. W., Lee, S. C., Chiang, H. Y., Syu, Y. C., Yu, X. X., and Hsieh, C. L. (2017). Psychometric properties of three measures assessing advanced theory of mind: evidence from people with schizophrenia. *Psychiatry Res.* 257, 490–496. doi: 10.1016/j.psychres.2017.08.026

Chiu, E. C., Wang, Y. C., Huang, S. L., Hsueh, I. P., Chiang, H. Y., and Hsieh, C. L. (2023). Test-retest reliabilities and minimal detectable changes of 5 versions of the Alzheimer's disease assessment scale-cognitive subscale in people with dementia. *Disabil. Rehabil.* 45, 1398–1404. doi: 10.1080/09638288.2022.2060334

Corcoran, R., Mercer, G., and Frith, C. D. (1995). Schizophrenia, symptomatology and social inference: investigating "theory of mind" in people with schizophrenia. *Schizophr. Res.* 17, 5–13. doi: 10.1016/0920-9964(95)00024-g

Cotter, J., Firth, J., Enzinger, C., Kontopantelis, E., Yung, A. R., Elliott, R., et al. (2016). Social cognition in multiple sclerosis: a systematic review and meta-analysis. *Neurology* 87, 1727–1736. https://10.1212/WNL.0000000000003236. doi: 10.1212/WNL.0000000000003236

d'Arma, A., Isernia, S., Di Tella, S., Rovaris, M., Valle, A., Baglio, F., et al. (2021). Social cognition training for enhancing affective and cognitive theory of mind in schizophrenia: a systematic review and a Meta-analysis. *J. Psychol.* 155, 26–58. doi: 10.1080/00223980.2020.1818671

d'Arma, A., Valle, A., Massaro, D., Baglio, G., Isernia, S., Di Tella, S., et al. (2023). A cultural training for the improvement of cognitive and affective theory of mind in people with multiple sclerosis: a pilot randomized controlled study. *Front. Psychol.* 14:1198018. doi: 10.3389/fpsyg.2023.1198018

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., and Bouter, L. M. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual. Life Outcomes* 4:54. doi: 10.1186/1477-7525-4-54

Frith, C. D. (2008). Social cognition. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 363, 2033–2039. doi: 10.1098/rstb.2008.0005

Happè, F. (2021). "Attributions (first order/second order)" in Encyclopedia of Autism Spectrum Disorders, 394–395.

Henry, J. D., von Hippel, W., Molenberghs, P., Lee, T., and Sachdev, P. S. (2016). Clinical assessment of social cognitive function in neurological disorders. *Nat. Rev. Neurol.* 12, 28–39. doi: 10.1038/nrneurol.2015.229

Isernia, S., Rossetto, F., Blasi, V., Massaro, D., Castelli, I., Ricci, C., et al. (2022a). Measuring cognitive and affective theory of mind with the Italian Yoni task: normative data and short versions. *Curr. Psychol.* 42, 23519–23530. doi: 10.1007/s12144-022-03457-5

Isernia, S., Rossetto, F., Shamay-Tsoory, S., Marchetti, A., and Baglio, F. (2022b). Standardization and normative data of the 48-item Yoni short version for the assessment of theory of mind in typical and atypical conditions. *Front. Aging Neurosci.* 14:1048599. doi: 10.3389/fnagi.2022.1048599

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., et al. (2010). Dissociating cognitive from affective theory of mind: a TMS study. *Cortex* 46, 769–780. doi: 10.1016/j.cortex.2009.07.010

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: the social attribution task. *J. Child Psychol. Psychiatry* 41, 831–846. doi: 10.1111/1469-7610.00671

Koo, B. M., and Vizer, L. M. (2019). Mobile Technology for Cognitive Assessment of older adults: a scoping review. Innov. *Aging* 3:igy038. doi: 10.1093/geroni/igy038

Lee, P., Liu, C. H., Fan, C. W., Lu, C. P., Lu, W. S., and Hsieh, C. L. (2013). The test-retest reliability and the minimal detectable change of the Purdue pegboard test in schizophrenia. *J. Formos. Med. Assoc.* 112, 332–337. doi: 10.1016/j.jfma.2012.02.023

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., et al. (1999). The TOM test: a new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *J. Autism Dev. Disord.* 29, 67–80. doi: 10.1023/a:1025922717020

Negrete, R., Simms, S., Gross, J., Nunes Rabello, L., Hixon, M., Zeini, I. M., et al. (2021). The test re-test reliability of a novel single leg hop test (T-Drill hop test). *Int. J. Sports Phys. Ther.* 16, 724–731. doi: 10.26603/001c.23677

Palmer, S., Manns, S., Cramp, F., Lewis, R., and Clark, E. M. (2017). Test-retest reliability and smallest detectable change of the Bristol impact of hypermobility (BIoH) questionnaire. *Musculoskelet. Sci. Pract.* 32, 64–69. doi: 10.1016/j.msksp.2017.08.007

Pirastru, A., Di Tella, S., Cazzoli, M., Esposito, F., Baselli, G., Baglio, F., et al. (2023). The impact of emotional valence and stimulus habituation on fMRI signal reliability during emotion generation. *NeuroImage* 284:120457. doi: 10.1016/j.neuroimage.2023.120457

Plana, I., Lavoie, M. A., Battaglia, M., and Achim, A. M. (2014). A meta-analysis and scoping review of social cognition performance in social phobia, posttraumatic stress disorder and other anxiety disorders. *J. Anxiety Disord.* 28, 169–177. doi: 10.1016/j.janxdis.2013.09.005

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/S0140525X00076512

Rosseel, Y. (2012). lavaan: a brief user's guide. Index of/yrosseel/lavaan.

Rossetto, F., Baglio, F., Massaro, D., Alberoni, M., Nemni, R., Marchetti, A., et al. (2020). Social cognition in rehabilitation context: different evolution of affective and cognitive theory of mind in mild cognitive impairment. *Behav. Neurol.* 2020, 5204927–5204929. doi: 10.1155/2020/5204927

Rossetto, F., Castelli, I., Baglio, F., Massaro, D., Alberoni, M., Nemni, R., et al. (2018). Cognitive and affective theory of mind in mild cognitive impairment and Parkinson's disease: preliminary evidence from the Italian version of the Yoni task. *Dev. Neuropsychol.* 43, 764–780. doi: 10.1080/87565641.2018.1529175

Rossetto, F., Isernia, S., Cabinio, M., Pirastru, A., Blasi, V., and Baglio, F. (2022). Affective theory of mind as a residual ability to preserve mentalizing in amnestic mild cognitive impairment: a 12-months longitudinal study. *Front. Neurol.* 13:1060699. doi: 10.3389/fneur.2022.1060699

Shamay-Tsoory, S. G., and Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia* 45, 3054–3067. doi: 10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Aharon-Peretz, J., and Levkovitz, Y. (2007). The neuroanatomical basis of affective mentalizing in schizophrenia: comparison of patients with schizophrenia and patients with localized prefrontal lesions. *Schizophr. Res.* 90, 274–283. doi: 10.1016/j.schres.2006.09.020

Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., and Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. *Cogn. Behav. Neurol.* 18, 55–67. doi: 10.1097/01.wnn.0000152228.90129.99

Stone, V. E., Baron-Cohen, S., Calder, A., Keane, J., and Young, A. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia* 41, 209–220. doi: 10.1016/s0028-3932(02)00151-3

Stone, V. E., Baron-Cohen, S., and Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *J. Cogn. Neurosci.* 10, 640–656. doi: 10.1162/089892998562942

Togher, L., Douglas, J., Turkstra, L. S., Welch-West, P., Janzen, S., Harnett, A., et al. (2023). INCOG 2.0 guidelines for cognitive rehabilitation following traumatic brain injury, part IV: cognitive-communication and social cognition disorders. *J. Head Trauma Rehabil.* 38, 65–82. doi: 10.1097/HTR.0000000000000835

Umberson, D., and Montez, J. K. (2010). Social relationships and health: a flashpoint for health policy. *J. Health Soc. Behav.* 51, S54–S66. doi: 10.1177/0022146510383501

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., et al. (2013). The "Reading the mind in the eyes" test: systematic review of psychometric properties and a validation study in Italy. *Cogn. Neuropsychiatry* 18, 326–354. doi: 10.1080/13546805.2012.721728

Watson, P. F., and Petrie, A. (2010). Method agreement analysis: a review of correct methodology. *Theriogenology* 73, 1167–1179. doi: 10.1016/j.theriogenology.2010.01.003

Webb, K. L., Ryan, J., Wolfe, R., Woods, R. L., Shah, R. C., Murray, A. M., et al. (2022). Test-retest reliability and minimal detectable change of four cognitive tests in community-dwelling older adults. *J. Alzheimers Dis.* 87, 1683–1693. doi: 10.3233/jad-215564

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Yeh, Y. C., Lin, C. Y., Li, P. C., Hung, C. F., Cheng, C. H., Kuo, M. H., et al. (2021). A systematic review of the current measures of theory of mind in adults with schizophrenia. *Int. J. Environ. Res. Public Health* 18:7172. doi: 10.3390/ijerph18137172

Yi, Z., Zhao, P., Zhang, H., Shi, Y., Shi, H., Zhong, J., et al. (2020). Theory of mind in Alzheimer's disease and amnestic mild cognitive impairment: a meta-analysis. *Neurol. Sci.* 41, 1027–1039. doi: 10.1007/s10072-019-04215-5

frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 13 August 2024
DOI 10.3389/fpsyg.2024.1452462

OPEN ACCESS

EDITED BY
Alessio Facchin,
Magna Graecia University, Italy

REVIEWED BY
Ruben Gur,
University of Pennsylvania, United States
Tyler M. Moore,
University of Pennsylvania, United States, in collaboration with reviewer RG
Rachael L. Ellison,
Rosalind Franklin University of Medicine and Science, United States

*CORRESPONDENCE
John-Christopher A. Finley
✉ jfinley3045@gmail.com

RECEIVED 20 June 2024
ACCEPTED 29 July 2024
PUBLISHED 13 August 2024

CITATION
Finley J-CA (2024) Performance validity testing: the need for digital technology and where to go from here.
Front. Psychol. 15:1452462.
doi: 10.3389/fpsyg.2024.1452462

Performance validity testing: the need for digital technology and where to go from here

John-Christopher A. Finley*

Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

KEYWORDS

performance validity, malinger, feign, digital, artificial intelligence, technology, neuropsychology, computerized

Introduction

Neuropsychological testing can inform practitioners and scientists about brain-behavior relationships that guide diagnostic classification and treatment planning (Donders, 2020). However, not all examinees remain engaged throughout testing and some may exaggerate or feign impairment, rendering their performance non-credible and uninterpretable (Roor et al., 2024). It is therefore important to regularly assess the validity of data obtained during a neuropsychological evaluation (Sweet et al., 2021). However, performance validity assessment (PVA) is a complex process. Practitioners must know when and how to use multiple performance validity tests (PVTs) while accounting for various contextual, diagnostic, and intrapersonal factors (Lippa, 2018). Furthermore, inaccurate PVA can lead to erroneous and potentially harmful judgments regarding an examinee's mental health and neuropsychological status. Although the methods used to address these complexities in PVA are evolving (Bianchini et al., 2001; Boone, 2021), improvement is still needed.

Modern digital technologies have the potential to significantly improve PVA, but such technologies have not received much attention. Most PVTs used today are pencil-and-paper tests developed several decades ago (Martin et al., 2015), and digital innovations have largely been confined to computerized validity testing (see Table 1). Meanwhile, other areas of digital neuropsychology have rapidly expanded. Technologies can now capture high-dimensional data conducive to precision medicine (Parsons and Duffield, 2020; Harris et al., 2024), and this surge in digital assessment may soon become the rule rather than exception for neuropsychology (Bilder and Reise, 2019; Germine et al., 2019). If PVA does not keep pace with other digital innovations in neuropsychology, many validity tests and methods may lose relevance.

This paper aims to increase awareness of how digital technologies can improve PVA so that researchers within neuropsychology and relevant organizations have a clinically and scientifically meaningful basis for transitioning to digital platforms. Herein, I describe five ways in which digital technologies can improve PVA: (1) generating more informative data, (2) leveraging advanced analytics, (3) facilitating scalable and sustainable research, (4) increasing accessibility, and (5) enhancing efficiencies.

Generating more informative data

Generating a greater volume, variety, and velocity of data core and ancillary to validity testing may improve the detection of non-credible performance. With these data, scientists and practitioners can better understand the dimensionality of performance validity and

assess it effectively, especially in cases without clear evidence of fabrication. However, capturing sundry data in PVA is challenging, as practitioners are often limited to a few PVTs throughout an evaluation that is completed in a single snapshot of time (Martin et al., 2015). Furthermore, many PVTs index redundant information because they have similar detection paradigms that generate only one summary cut-score (Boone, 2021). Digital technologies can address these issues by capturing additional aspects of performance validity without increasing time or effort.

Digitally recording the testing process is one way to generate more diverse data points than a summary score. Some process-based metrics are already employed in PVA, including recording response consistency and exaggeration across test items (Schroeder et al., 2012; Finley et al., 2024a). For example, Leese et al. (2024a) found that using a digital software to assess discrepancies between item responses and correct answers improved the detection of non-credible performance. Using digital tools to objectively and unobtrusively record response latencies and reaction times during testing is another useful process-based approach (Erdodi and Lichtenstein, 2021; Rhoads et al., 2021). Examinees typically cannot maintain consistent rates of slowed response latencies across items when attempting to feign impairment (Gutiérrez and Gur, 2011). Various software can record these process-based scores (e.g., item-level indices of response time, reliable span, and exaggeration magnitude) in most existing tests if they are migrated to tablets/computers (Kush et al., 2012). Recording both the process and outcome (summary scores) of test completion can index dimensions of performance validity across and within tests.

Technologies can also record biometric data ancillary to validity testing. Biometrics including oculomotor, cardiovascular, body gesture, and electrodermal responses are indicators of cognitive load and are associated with deception (Ayres et al., 2021). Deception is believed to increase cognitive load because it requires more complex processing to falsify a response (Dinges et al., 2024). Although deception is different from non-credible performance, neuroimaging research suggests non-credible performance can be indicative of greater cognitive effort (Allen et al., 2007). For this reason, technologies like eye-tracking have been used to augment PVA (Braw et al., 2024). These studies are promising, but other avenues within this literature have yet to be explored due to technological limitations. Fortunately, many technologies now possess built-in cameras, accelerometers, gyroscopes, and sensors that "see," "hear," and "feel" at a basic level, and may be embedded within existing PVTs to record biometrics.

Technologies under development for cognitive testing may also provide informative data that has not yet been linked to PVA. For example, speech analysis software for verbal fluency tasks (Holmlund et al., 2019) could identify non-credible word choice or grammatical errors. Similarly, digital phenotyping technologies may identify novel and useful indices during validity testing, such as keystroke dynamics (e.g., slowed/inconsistent typing; Chen et al., 2022) embedded with PVTs requiring typed responses. These are among many burgeoning technologies that can generate higher dimensional data needed for robust PVA without adding time or labor. However, access to a greater range and depth of data requires advanced methods to effectively and efficiently analyze the data.

Leveraging advanced analytics

Fortunately, technologies can leverage advanced analytics to rapidly and accurately analyze a large influx of digital data in real time. Although several statistical approaches are described within the PVA literature (Boone, 2021; Jewsbury, 2023), machine learning (ML) and item response theory (IRT) analytics may be particularly useful for analyzing large volumes of interrelated, nonlinear, and high-dimensional data at the item level (Reise and Waller, 2009; Mohri et al., 2012).

Not only can these approaches analyze more complex data but they can also improve the development and refinement of PVTs relative to classical measurement approaches. For example, person-fit statistics is an IRT approach that has been used to identify non-credible symptom reporting in dichotomous and polytomous data (Beck et al., 2019). This approach may also improve embedded PVTs by estimating the extent to which each item-level response deviates from one's true abilities (Bilder and Reise, 2019). Scott et al. (2023) found that using person-fit statistics helped embedded PVTs detect subtle patterns of non-credible performance. IRT is especially amenable to computerized adaptive testing, which adjusts each item's difficulty based on one's response. Computerized adaptive testing systems can create shorter and more precise PVTs with psychometrically equivalent alternative forms (Gibbons et al., 2008). These systems can also detect careless responding based on unpredictable error patterns that deviate from normal difficulty curves. Detecting careless responding may be useful for PVTs embedded within digital self-paced continuous performance tests (e.g., Nicholls et al., 2020; Berger et al., 2021). Other IRT approaches can improve PVTs by scrutinizing item difficulty and discriminatory power and identifying culturally biased items. For example, differential item functioning is an IRT approach that may identify items on English-verbally mediated PVTs that are disproportionately challenging for those who do not speak English as their primary language, allowing for appropriate adjustments.

ML has proven useful in symptom validity test development (Orrù et al., 2021) and may function similarly for PVTs. Two studies recently investigated whether supervised ML improves PVA (Pace et al., 2019; Hirsch et al., 2022). Pace et al. (2019) found that a supervised ML model trained with various features (demographics, cognitive performance errors, response time, and a PVT score) discriminated between genuine and simulated cognitive impairment with high accuracy. Using similar features, Hirsch et al. (2022) found that their supervised models had moderate to weak prediction of PVT failure in a clinical attention-deficit/hyperactivity disorder sample. No studies have used unsupervised ML for PVA. It is possible that unsupervised ML could also identify groups of credible and non-credible performing examinees using relevant factors such as PVT scores, litigation status, medical history, and referral reasons, without explicit programming. Software can be developed to extract data for the ML via computerized questionnaires or electronic medical records. Deep learning, a form of ML that processes data using multiple dimensions, may also detect complex and anomalous patterns indicative of non-credible performance. Deep learning may be especially

TABLE 1 Existing digital performance validity tests and methods.

Material-specificity	Performance validity test/method	References
Memory-focused freestanding PVTs	Memory integrated language test (MIL)	Finley et al., 2024b; Leese et al., 2024b
	Coin in hand–extended version	Daugherty et al., 2021
	Inventory of problems – memory (IOP-M)	Giromini et al., 2020; Erdodi et al., 2024
	DETECTS	Paulo and Albuquerque, 2019
	Computerized forced-choice test (CFCT)	Gutiérrez and Gur, 2011
	Medical symptom validity test (MSVT)	Green, 2004
	Word memory test (WMT)	Green, 2003
	Computerized test of memory malingering (TOMM)	Rees et al., 1998
	Computerized assessment of response bias (CARB)	Allen et al., 1997
	Tests of neuropsychological malingering (TNM)	Pritchard and Moses, 1992
Non-memory-focused freestanding PVTs	Making change test (MCT)	Finley et al., 2024b; Leese et al., 2024a
	The shell game task*	Bryant et al., 2023
	Multi-level pattern memory test (MPMT)	Omer and Braw, 2021
	Tests of attentional distraction (TOAD)	Morey, 2019
	Nonverbal medical symptom validity test (NV-MSVT)	Green, 2008
	Portland digit recognition test-computerized	Rose et al., 1995
	Victoria symptom validity test (VSVT)	Slick et al., 1995
	Forced choice test of nonverbal ability (FCTNV)	Frederick and Foster, 1991
	Multi-digit memory test (MDMT)	Bolter and Niccolls, 1991
Mixed freestanding PVTs	Pediatric performance validity test suite (PdPVTS)	McCaffrey et al., 2020
	Memory validity profile (MVP)	Brooks and Sherman, 2019; Brooks et al., 2019
Embedded PVTs/methods	Penn computerized neurocognitive battery (PennCNB)	Scott et al., 2023
	National Institutes of Health Toolbox® (NIHTB)	Abeare et al., 2021
	MOXO-d-continuous performance test (CPT)	Berger et al., 2021; Winter and Braw, 2022
	Conners continuous performance test (CPT; Versions 2 and 3)	Ord et al., 2010; Erdodi et al., 2014; Shura et al., 2016; Sharland et al., 2018; Lichtenstein et al., 2019; Scimeca et al., 2021; Finley et al., 2023a,b; Robinson et al., 2023;
	Test of variables of attention (TOVA)	Leark et al., 2002; Marshall et al., 2010; Nicholls et al., 2020
	Automated neuropsychological assessment metrics (ANAM) performance validity index	Roebuck-Spencer et al., 2013; Meyers et al., 2022
	Immediate post-concussion assessment and cognitive testing (ImPACT)	Erdal, 2012; Schatz and Glatts, 2013; Lovell, 2015; Siedlik et al., 2015; Gaudet and Weyandt, 2017; Higgins et al., 2017; Manderino and Gunstad, 2018; Raab et al., 2020
	CNS vital signs battery	Brooks et al., 2014
	NeuroTrax battery	Hegedish et al., 2012; Bar-Hen et al., 2015

*Presented as a professional conference poster, not a published article.

useful for analyzing response sequences over time (e.g., non-credible changes in performance across repeat medico-legal evaluations). Furthermore, deep-learning models may be effective at identifying inherent statistical dependencies and patterns of non-credible performance, and thus generating expectations of how genuine responses should appear. Combining these algorithms with other statistical techniques that assess response complexity and highly anomalous responses (e.g., Lundberg and Lee, 2017; Parente and Finley, 2018; Finley and Parente, 2020; Orrù et al., 2020; Mertler et al., 2021; Parente et al., 2021, 2023; Finley et al., 2022; Rodriguez et al., 2024), may increase the signal of non-credible performance. These algorithmic approaches can improve as we better understand cognitive phenotypes and what is improbable for certain disorders using precision medicine and bioinformatics.

Facilitating scale and sustainability

To optimize the utility of these digital data, technologies can include point-of-testing acquisition software that automatically

transfers data to cloud-based, centralized repositories. These repositories facilitate sustainable and scalable innovations by increasing data access and collaboration among PVA stakeholders (see Reeves et al., 2007 and Gaudet and Weyandt, 2017 for large-scale developments of digital tests with embedded PVTs). Multidisciplinary approaches are needed to make theoretical and empirical sense of the data collected via digital technologies (Collins and Riley, 2016). With more comprehensive and uniform data amenable to data mining and deep-learning analytics, collaborating researchers can address overarching issues that remain poorly understood within research. For example, with larger centralized data researchers can directly evaluate different statistical approaches (e.g., chaining likelihood ratios vs. multivariable discriminant function analysis, Bayesian model averaging, or logistic regression) as well as the joint validity of standardized test batteries (Davis, 2021; Erdodi, 2023; Jewsbury, 2023). Such data and findings could also help determine robust criterion-grouping combinations, given that multiple PVTs assessing complementary aspects of performance across various cognitive domains may be necessary for a strong criterion-grouping combination (Schroeder et al., 2019; Soble et al., 2020). Similarly, researchers could expand upon existing decision-making models (e.g., Rickards et al., 2018; Sherman et al., 2020) by using these comprehensive data to develop algorithms that automatically generate credible/non-credible profiles based on the type and proportion or number of PVTs failed in relation to various contextual and diagnostic factors, symptom presentations, and clinical inconsistencies (across medical records, self- and informant-reports, or behavioral observations). A greater range and depth of data may further help elucidate the extent to which several putative factors—such as bona fide injury/disease, normal fluctuation and variability in testing, level of effort (either to perform well or to deceive), and symptom validity, among others—are associated with performance validity (Larrabee, 2012; Bigler, 2014). Understanding these associations could help identify the mechanisms underlying non-credible performance.

Collaboration is especially needed for basic and applied sciences to coalesce unique aspects of PVA that have been studied independently, such as integrating neuropsychology and neurocognitive processing theories to develop more sophisticated stimuli/paradigms (Leighton et al., 2014). For example, less applied scientific models, such as memory familiarity vs. conscious recollection theories, may be applied to clinically available PVTs to reduce false-positive rates in certain neurological populations (Eglit et al., 2017). Similar areas of cognitive science have also shown that using pictorial or numerical stimuli (vs. words) across multiple learning trials can reduce false-positive errors in clinical settings (Leighton et al., 2014). Furthermore, integrating data in real time into these repositories offers a sustainable and accurate way of estimating PVT failure base rates and developing cutoffs accordingly. Finally, as proposed by the National Neuropsychology Network (Loring et al., 2022), a centralized repository for digital data that is backward-compatible with analog test data can provide a smooth transition from traditional pencil-and-paper tests to digital formats. These repositories (including those curated via the National Neuropsychology Network) thus enable sustainable innovation by supporting continuous incremental refinement of PVTs over time.

Increasing accessibility

As observed in other areas of neuropsychology (Miller and Barr, 2017), digital technologies can offer more accessible PVA. Specifically, web-based PVTs can help access underserved and geographically restricted communities, but with the understanding that disparities in digital technology may also exist. Although more web-based PVTs are needed, not every PVT requires digitization for telehealth (e.g., Reliable Digit Span; Kanser et al., 2021; Harrison and Davin, 2023). Digital PVTs can also increase accessibility in primary care settings where digital cognitive screeners are being developed for face-to-face evaluations and may be completed in distracting, unsupervised environments (Zygouris and Tsolaki, 2015). Validity indicators could be embedded within these screeners rather than creating new freestanding PVTs. The National Institutes of Health Toolbox® (Abeare et al., 2021) and Penn Computerized Neurocognitive Battery (Scott et al., 2023) are well-established digital screeners with embedded PVTs that offer great promise for these evaluations. In primary care, embedded PVTs could serve as preliminary screeners for atypical performance that warrants further investigation. Digital PVTs may also increase accessibility in research settings. Although it is not highly likely research volunteers would deliberately feign impairment, they may lose interest, doze off, or rush through testing (An et al., 2017), especially in dementia-focused research where digital testing is common. Some digitally embedded PVTs have been developed for ADHD research (Table 1) and may be used in other research focused on digital cognitive testing (Bauer et al., 2012).

Enhancing efficiencies

Finally, the application of digital technologies introduces new efficiencies; in PVA, they hold the promise of improved standardization and administration/scoring accuracy. Technologies can leverage automated algorithms to reduce time spent on scoring and routine aspects of PVA (e.g., finding/adjusting PVT cutoffs according to various contextual/intrapersonal factors). Automation would allow providers to allocate more time to case conceptualization and responding to (rather than detecting) validity issues. Greater efficiencies in PVA translate into greater cost-efficiencies as well as reduced collateral expenses for specialized training, testing support, and materials (Davis, 2023). Further, digital PVTs can automatically store, retrieve, and analyze data to generate multiple relevant scores (e.g., specificity, sensitivity, predictive power adjusted for diagnostic-specific base rates, false-positive estimates, and likelihood ratios or probability estimates for single/multivariable failure combinations). Automated scoring will likely become increasingly useful as more PVTs and data are generated.

Limitations and concluding remarks

By no means an exhaustive review, this paper describes five ways in which digital technologies can improve PVA. These

improvements can complement rather than replace the uniquely human aspects of PVA. Thus, the upfront investments required to transition to digital approaches are likely justifiable. However, other limitations deserve attention before making this transition. As described elsewhere (Miller and Barr, 2017; Germine et al., 2019), limitations to digital assessment may include variability across devices, which can impose different perceptual, motor, and cognitive demands that affect the reliability and accuracy of the tests. Variations in hardware and software within the same class of devices can affect stimulus presentation and response (including response latency) measurement. Individual differences in access to and familiarity with technology may further affect test performance. Additionally, the rapid advancement in technologies suggests that hardware and software can quickly become obsolete. A large influx of data and the application of "black box" ML algorithms and cloud-based repositories also raises concerns regarding data security and privacy. Addressing these issues and implementing digital methods into practice or research would require substantial technological and human infrastructure that may not be attainable in certain settings (Miller, 2019). Indeed, the utility of digital assessments likely depends on the context in which they are implemented. For example, PVA is critical in forensic evaluations but the limitations described above could challenge compliance with the evolving standards for the admissibility of scientific evidence in these evaluations. Further discussion of these limitations along with the logistical and practical considerations for a digital transition is needed (for further discussion, see Miller, 2019; Singh and Germine, 2021). Finally, other digital opportunities, such as using validity indicators with ecological momentary assessment and virtual reality technologies, merit further discussion. Moving forward, scientists are encouraged to expand upon these digital innovations to ensure that PVA evolves alongside the broader landscape of digital neuropsychology.

Author contributions

J-CF: Conceptualization, Investigation, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

I would like to thank Jason Soble and Anthony Robinson for providing their expertise and guidance during the preparation of this manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abeare, C., Erdodi, L., Messa, I., Terry, D. P., Panenka, W. J., Iverson, G. L., et al. (2021). Development of embedded performance validity indicators in the NIH Toolbox Cognitive Battery. *Psychol. Assess.* 33, 90–96. doi: 10.1037/pas0000958

Allen, L. M., Conder, R. L., Green, P., and Cox, D. R. (1997). *CARB'97 manual for the computerized assessment of response bias.* Durham, NC: CogniSyst.

Allen, M. D., Bigler, E. D., Larsen, J., Goodrich-Hunsaker, N. J., and Hopkins, R. O. (2007). Functional neuroimaging evidence for high cognitive effort on the Word Memory Test in the absence of external incentives. *Brain Injury* 21, 1425–1428. doi: 10.1080/02699050701769819

An, K. Y., Kaploun, K., Erdodi, L. A., and Abeare, C. A. (2017). Performance validity in undergraduate research participants: A comparison of failure rates across tests and cutoffs. *Clin. Neuropsychol.* 31, 193–206. doi: 10.1080/13854046.2016.1217046

Ayres, P., Lee, J. Y., Paas, F., and Van Merrienboer, J. J. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Front. Psychol.* 12:702538. doi: 10.3389/fpsyg.2021.702538

Bar-Hen, M., Doniger, G. M., Golzad, M., Geva, N., and Schweiger, A. (2015). Empirically derived algorithm for performance validity assessment embedded in a widely used neuropsychological battery: validation among TBI patients in litigation. *J. Clin. Exper. Neuropsychol.* 37, 1086–1097. doi: 10.1080/13803395.2015.1078294

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., and Naugle, R. I. (2012). Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch. Clin. Neuropsychol.* 27, 362–373. doi: 10.1093/arclin/acs027

Beck, M. F., Albano, A. D., and Smith, W. M. (2019). Person-fit as an index of inattentive responding: a comparison of methods using polytomous survey data. *Appl. Psychol. Meas.* 43, 374–387. doi: 10.1177/0146621618798666

Berger, C., Lev, A., Braw, Y., Elbaum, T., Wagner, M., and Rassovsky, Y. (2021). Detection of feigned ADHD using the MOXO-d-CPT. *J. Atten. Disord.* 25, 1032–1047. doi: 10.1177/1087054719864656

Bianchini, K. J., Mathias, C. W., and Greve, K. W. (2001). Symptom validity testing: a critical review. *Clin. Neuropsychol.* 15, 19–45. doi: 10.1076/clin.15.1.19.1907

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain Injury* 28, 1623–1638. doi: 10.3109/02699052.2014.947627

Bilder, R. M., and Reise, S. P. (2019). Neuropsychological tests of the future: How do we get there from here?. *Clin. Neuropsychol.* 33, 220–245. doi: 10.1080/13854046.2018.1521993

Bolter, J. F., and Niccolls, R. (1991). *Multi-Digit Memory Test.* Wang Neuropsychological Laboratories. Boone, K. B. (2021). *Assessment of Feigned Cognitive Impairment*. London: Guilford Publications.

Boone, K. B. (2021). *Assessment of Feigned Cognitive Impairment*. London: Guilford Publications.

Braw, Y. C., Elbaum, T., Lupu, T., and Ratmansky, M. (2024). Chronic pain: Utility of an eye-tracker integrated stand-alone performance validity test. *Psychol. Injury Law.* 13, 139–151. doi: 10.1007/s12207-024-09507-6

Brooks, B. L., Fay-McClymont, T. B., MacAllister, W. S., Vasserman, M., and Sherman, E. M. (2019). A new kid on the block: the memory validity profile (MVP) in children with neurological conditions. *Child Neuropsychol.* 25, 561–572. doi: 10.1080/09297049.2018.1477929

Brooks, B. L., and Sherman, E. M. (2019). Using the Memory Validity Profile (MVP) to detect invalid performance in youth with mild traumatic brain injury. *Appl. Neuropsychol.* 8, 319–325. doi: 10.1080/21622965.2018.1476865

Brooks, B. L., Sherman, E. M., and Iverson, G. L. (2014). Embedded validity indicators on CNS Vital Signs in youth with neurological diagnoses. *Arch. Clin. Neuropsychol.* 29, 422–431. doi: 10.1093/arclin/acu029

Bryant, A. M., Pizzonia, K., Alexander, C., Lee, G., Revels-Strother, O., Weekman, S., et al. (2023). 77 The Shell Game Task: Pilot data using a simulator-design study to evaluate a novel attentional performance validity test. *J. Int. Neuropsychol. Soc.* 29, 751–752. doi: 10.1017/S1355617723009359

Chen, M. H., Leow, A., Ross, M. K., DeLuca, J., Chiaravalloti, N., Costa, S. L., et al. (2022). Associations between smartphone keystroke dynamics and cognition in MS. *Digital Health* 8:234. doi: 10.1177/20552076221143234

Collins, F. S., and Riley, W. T. (2016). NIH's transformative opportunities for the behavioral and social sciences. *Sci. Transl. Med.* 8, 366ed14. doi: 10.1126/scitranslmed.aai9374

Daugherty, J. C., Querido, L., Quiroz, N., Wang, D., Hidalgo-Ruzzante, N., Fernandes, S., et al. (2021). The coin in hand–extended version: development and validation of a multicultural performance validity test. *Assessment* 28, 186–198. doi: 10.1177/1073191119864652

Davis, J. J. (2021). "Interpretation of data from multiple performance validity tests," in *Assessment of feigned cognitive impairment,* ed. K. B. Boone (London: Guilford Publications), 283–306.

Davis, J. J. (2023). Time is money: Examining the time cost and associated charges of common performance validity tests. *Clin. Neuropsychol.* 37, 475–490. doi: 10.1080/13854046.2022.2063190

Dinges, L., Fiedler, M. A., Al-Hamadi, A., Hempel, T., Abdelrahman, A., Weimann, J., et al. (2024). Exploring facial cues: automated deception detection using artificial intelligence. *Neural Comput. Applic.* 26, 1–27. doi: 10.1007/s00521-024-09811-x

Donders, J. (2020). The incremental value of neuropsychological assessment: a critical review. *Clin. Neuropsychol.* 34, 56–87. doi: 10.1080/13854046.2019.1575471

Eglit, G. M., Lynch, J. K., and McCaffrey, R. J. (2017). Not all performance validity tests are created equal: the role of recollection and familiarity in the Test of Memory Malingering and Word Memory Test. *J. Clin. Exp. Neuropsychol.* 39, 173–189. doi: 10.1080/13803395.2016.1210573

Erdal, K. (2012). Neuropsychological testing for sports-related concussion: how athletes can sandbag their baseline testing without detection. *Arch. Clin. Neuropsychol.* 27, 473–479. doi: 10.1093/arclin/acs050

Erdodi, L., Calamia, M., Holcomb, M., Robinson, A., Rasmussen, L., and Bianchini, K. (2024). M is for performance validity: The iop-m provides a cost-effective measure of the credibility of memory deficits during neuropsychological evaluations. *J. Forensic Psychol. Res. Pract.* 24, 434–450. doi: 10.1080/24732850.2023.2168581

Erdodi, L. A. (2023). Cutoff elasticity in multivariate models of performance validity assessment as a function of the number of components and aggregation method. *Psychol. Inj. Law* 16, 328–350. doi: 10.1007/s12207-023-09490-4

Erdodi, L. A., and Lichtenstein, J. D. (2021). Invalid before impaired: An emerging paradox of embedded validity indicators. *Clin. Neuropsychol.* 31, 1029–1046. doi: 10.1080/13854046.2017.1323119

Erdodi, L. A., Roth, R. M., Kirsch, N. L., Lajiness-O'Neill, R., and Medoff, B. (2014). Aggregating validity indicators embedded in Conners' CPT-II outperforms individual cutoffs at separating valid from invalid performance in adults with traumatic brain injury. *Arch. Clin. Neuropsychol.* 29, 456–466. doi: 10.1093/arclin/acu026

Finley, J. C. A., Brook, M., Kern, D., Reilly, J., and Hanlon, R. (2023b). Profile of embedded validity indicators in criminal defendants with verified valid neuropsychological test performance. *Arch. Clin. Neuropsychol.* 38, 513–524. doi: 10.1093/arclin/acac073

Finley, J. C. A., Brooks, J. M., Nili, A. N., Oh, A., VanLandingham, H. B., Ovsiew, G. P., et al. (2023a). Multivariate examination of embedded indicators of performance validity for ADHD evaluations: a targeted approach. *Appl. Neuropsychol.* 23, 1–17. doi: 10.1080/23279095.2023.2256440

Finley, J. C. A., Kaddis, L., and Parente, F. J. (2022). Measuring subjective clustering of verbal information after moderate-severe traumatic brain injury: A preliminary review. *Brain Injury* 36, 1019–1024. doi: 10.1080/02699052.2022.2109751

Finley, J. C. A., Leese, M. I., Roseberry, J. E., and Hill, S. K. (2024b). Multivariable utility of the Memory Integrated Language and Making Change Test. *Appl. Neuropsychol. Adult* 1–8. doi: 10.1080/23279095.2024.2385439

Finley, J. C. A., and Parente, F. J. (2020). Organization and recall of visual stimuli after traumatic brain injury. *Brain Injury* 34, 751–756. doi: 10.1080/02699052.2020.1753113

Finley, J. C. A., Rodriguez, C., Cerny, B., Chang, F., Brooks, J., Ovsiew, G., et al. (2024a). Comparing embedded performance validity indicators within the WAIS-IV Letter-Number Sequencing subtest to Reliable Digit Span among adults referred for evaluation of attention deficit/hyperactivity disorder. *Clin. Neuropsychol.* 2024, 1–17. doi: 10.1080/13854046.2024.2315738

Frederick, R. I., and Foster, H. G. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychol. Assess.* 3, 596–602. doi: 10.1037/1040-3590.3.4.596

Gaudet, C. E., and Weyandt, L. L. (2017). Immediate Post-Concussion and Cognitive Testing (ImPACT): a systematic review of the prevalence and assessment of invalid performance. *Clin. Neuropsychol.* 31, 43–58. doi: 10.1080/13854046.2016.1220622

Germine, L., Reinecke, K., and Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *Clin. Neuropsychol.* 33, 271–286. doi: 10.1080/13854046.2018.1535662

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr. Serv.* 59, 361–368. doi: 10.1176/ps.2008.59.4.361

Giromini, L., Viglione, D. J., Zennaro, A., Maffei, A., and Erdodi, L. A. (2020). SVT Meets PVT: development and initial validation of the inventory of problems–memory (IOP-M). *Psychol. Inj. Law* 13, 261–274. doi: 10.1007/s12207-020-09385-8

Green, P. (2003). *Manual for the Word Memory Test for Windows.* Kelowna: Green's Publishing.

Green, P. (2004). *Green's Medical Symptom Validity Test (MSVT) for microsoft windows: User's manual*. Kelowna: Green's Publishing.

Green, P. (2008). *Green's Nonverbal Medical Symptom Validity Test (NV-MSVT) for microsoft windows: User's manual 1.0*. Kelowna: Green's Publishing.

Gutiérrez, J. M., and Gur, R. C. (2011). "Detection of malingering using forced-choice techniques," in *Detection of malingering during head injury litigation,* ed. C. R. Reynolds (Cham: Springer), 151–167. doi: 10.1007/978-1-4614-0442-2_4

Harris, C., Tang, Y., Birnbaum, E., Cherian, C., Mendhe, D., and Chen, M. H. (2024). Digital neuropsychology beyond computerized cognitive assessment: Applications of novel digital technologies. *Arch. Clin. Neuropsychol.* 39, 290–304. doi: 10.1093/arclin/acae016

Harrison, A. G., and Davin, N. (2023). Detecting non-credible performance during virtual testing. *Psychol. Inj. Law* 16, 264–272. doi: 10.1007/s12207-023-09480-6

Hegedish, O., Doniger, G. M., and Schweiger, A. (2012). Detecting response bias on the MindStreams battery. *Psychiat. Psychol. Law* 19, 262–281. doi: 10.1080/13218719.2011.561767

Higgins, K. L., Denney, R. L., and Maerlender, A. (2017). Sandbagging on the immediate post-concussion assessment and cognitive testing (ImPACT) in a high school athlete population. *Arch. Clin. Neuropsychol.* 32, 259–266. doi: 10.1093/arclin/acw108

Hirsch, O., Fuermaier, A. B., Tucha, O., Albrecht, B., Chavanon, M. L., and Christiansen, H. (2022). Symptom and performance validity in samples of adults at clinical evaluation of ADHD: a replication study using machine learning algorithms. *J. Clin. Exp. Neuropsychol.* 44, 171–184. doi: 10.1080/13803395.2022.2105821

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S., and Elvevåg, B. (2019). Updating verbal fluency analysis for the 21st century: applications for psychiatry. *Psychiatry Res.* 273, 767–769. doi: 10.1016/j.psychres.2019.02.014

Jewsbury, P. A. (2023). Invited commentary: Bayesian inference with multiple tests. *Neuropsychol. Rev.* 33, 643–652. doi: 10.1007/s11065-023-09604-4

Kanser, R. J., O'Rourke, J. J. F., and Silva, M. A. (2021). Performance validity testing via telehealth and failure rate in veterans with moderate-to-severe traumatic brain injury: a veterans affairs TBI model systems study. *NeuroRehabilitation* 49, 169–177. doi: 10.3233/NRE-218019

Kush, J. C., Spring, M. B., and Barkand, J. (2012). Advances in the assessment of cognitive skills using computer-based measurement. *Behav. Res. Methods* 44, 125–134. doi: 10.3758/s13428-011-0136-2

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *J. Int. Neuropsychol. Soc.* 18, 625–630. doi: 10.1017/S1355617712000240

Leark, R. A., Dixon, D., Hoffman, T., and Huynh, D. (2002). Fake bad test response bias effects on the test of variables of attention. *Arch. Clin. Neuropsychol.* 17, 335–342. doi: 10.1093/arclin/17.4.335

Leese, M. I., Finley, J. C. A., Roseberry, S., and Hill, S. K. (2024a). The Making Change Test: Initial validation of a novel digitized performance validity test for tele-neuropsychology. *Clin. Neuropsychol.* 2024, 1–14. doi: 10.1080/13854046.2024.2352898

Leese, M. I., Roseberry, J. E., Soble, J. R., and Hill, S. K. (2024b). The Memory Integrated Language Test (MIL test): initial validation of a novel web-based performance validity test. *Psychol. Inj. Law* 17, 34–44. doi: 10.1007/s12207-023-09495-z

Leighton, A., Weinborn, M., and Maybery, M. (2014). Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: a review and methodological approach. *J. Int. Neuropsychol. Soc.* 20, 873–886. doi: 10.1017/S135561771400085X

Lichtenstein, J. D., Flaro, L., Baldwin, F. S., Rai, J., and Erdodi, L. A. (2019). Further evidence for embedded performance validity tests in children within the Conners' continuous performance test–second edition. *Dev. Neuropsychol.* 44, 159–171. doi: 10.1080/87565641.2019.1565535

Lippa, S. M. (2018). Performance validity testing in neuropsychology: A clinical guide, critical review, and update on a rapidly evolving literature. *Clin. Neuropsychol.* 32, 391–421. doi: 10.1080/13854046.2017.1406146

Loring, D. W., Bauer, R. M., Cavanagh, L., Drane, D. L., Enriquez, K. D., Reise, S. P., et al. (2022). Rationale and design of the national neuropsychology network. *J. Int. Neuropsychol. Soc.* 28, 1–11. doi: 10.1017/S1355617721000199

Lovell (2015). *ImPACT test administration and interpretation manual*. Available at: http://www.impacttest.com (accessed July 23, 2024).

Lundberg, S. M., and Lee, S. I. (2017). "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, eds. I. Guyon, Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., et al. (New York: Curran Associates), 4765–4774.

Manderino, L., and Gunstad, J. (2018). Collegiate student athletes with history of ADHD or academic difficulties are more likely to produce an invalid protocol on baseline impact testing. *Clin. J. Sport Med.* 28, 111–116. doi: 10.1097/JSM.0000000000000433

Marshall, P., Schroeder, R., O'Brien, J., Fischer, R., Ries, A., Blesi, B., et al. (2010). Effectiveness of symptom validity measures in identifying cognitive and behavioral symptom exaggeration in adult attention deficit hyperactivity disorder. *Clin. Neuropsychol.* 24, 1204–1237. doi: 10.1080/13854046.2010.514290

Martin, P. K., Schroeder, R. W., and Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: a survey of North American professionals. *Clin. Neuropsychol.* 29, 741–776. doi: 10.1080/13854046.2015.1087597

McCaffrey, R. J., Lynch, J. K., Leark, R. A., and Reynolds, C. R. (2020). *Pediatric performance validity test suite (PdPVTS): Technical manual*. Multi-Health Systems, Inc.

Mertler, C. A., Vannatta, R. A., and LaVenia, K. N. (2021). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation*. London: Routledge. doi: 10.4324/9781003047223

Meyers, J. E., Miller, R. M., and Vincent, A. S. (2022). A validity measure for the automated neuropsychological assessment metrics. *Arch. Clin. Neuropsychol.* 37, 1765–1771. doi: 10.1093/arclin/acac046

Miller, J. B. (2019). Big data and biomedical informatics: Preparing for the modernization of clinical neuropsychology. *Clin. Neuropsychol.* 33, 287–304. doi: 10.1080/13854046.2018.1523466

Miller, J. B., and Barr, W. B. (2017). The technology crisis in neuropsychology. *Arch. Clin. Neuropsychol.* 32, 541–554. doi: 10.1093/arclin/acx050

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. London: The MIT Press.

Morey, L. C. (2019). Examining a novel performance validity task for the detection of feigned attentional problems. *Appl. Neuropsychol.* 26, 255–267. doi: 10.1080/23279095.2017.1409749

Nicholls, C. J., Winstone, L. K., DiVirgilio, E. K., and Foley, M. B. (2020). Test of variables of attention performance among ADHD children with credible vs. non-credible PVT performance. *Appl. Neuropsychol.* 9, 307–313. doi: 10.1080/21622965.2020.1751787

Omer, E., and Braw, Y. (2021). The Multi-Level Pattern Memory Test (MPMT): Initial validation of a novel performance validity test. *Brain Sci.* 11, 1039–1055. doi: 10.3390/brainsci11081039

Ord, J. S., Boettcher, A. C., Greve, K. W., and Bianchini, K. J. (2010). Detection of malingering in mild traumatic brain injury with the Conners' Continuous Performance Test–II. *J. Clin. Exp. Neuropsychol.* 32, 380–387. doi: 10.1080/13803390903066881

Orrù, G., Mazza, C., Monaro, M., Ferracuti, S., Sartori, G., and Roma, P. (2021). The development of a short version of the SIMS using machine learning to detect feigning in forensic assessment. *Psychol. Inj. Law* 14, 46–57. doi: 10.1007/s12207-020-09389-4

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Front. Psychol.* 10:2970. doi: 10.3389/fpsyg.2019.02970

Pace, G., Orrù, G., Monaro, M., Gnoato, F., Vitaliani, R., Boone, K. B., et al. (2019). Malingering detection of cognitive impairment with the B test is boosted using machine learning. *Front. Psychol.* 10:1650. doi: 10.3389/fpsyg.2019.01650

Parente, F. J., and Finley, J. C. A. (2018). Using association rules to measure subjective organization after acquired brain injury. *NeuroRehabilitation* 42, 9–15. doi: 10.3233/NRE-172227

Parente, F. J., Finley, J. C. A., and Magalis, C. (2021). An association rule general analytical system (ARGAS) for hypothesis testing in qualitative and quantitative research. *Int. J. Quant. Qualit. Res. Methods* 9, 1–13. Available online at: https://ssrn.com/abstract=3773480

Parente, F. J., Finley, J. C. A., and Magalis, C. (2023). A quantitative analysis for non-numeric data. *Int. J. Quant. Qualit. Res. Methods* 11, 1–11. doi: 10.37745/ijqqrm13/vol11n1111

Parsons, T., and Duffield, T. (2020). Paradigm shift toward digital neuropsychology and high-dimensional neuropsychological assessments. *J. Med. Internet Res.* 22:e23777. doi: 10.2196/23777

Paulo, R., and Albuquerque, P. B. (2019). Detecting memory performance validity with DETECTS: a computerized performance validity test. *Appl. Neuropsychol.* 26, 48–57. doi: 10.1080/23279095.2017.1359179

Pritchard, D., and Moses, J. (1992). Tests of neuropsychological malingering. *Forensic Rep.* 5, 287–290.

Raab, C. A., Peak, A. S., and Knoderer, C. (2020). Half of purposeful baseline sandbaggers undetected by ImPACT's embedded invalidity indicators. *Arch. Clin. Neuropsychol.* 35, 283–290. doi: 10.1093/arclin/acz001

Rees, L. M., Tombaugh, T. N., Gansler, D. A., and Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychol. Assess.* 10, 10–20. doi: 10.1037/1040-3590.10.1.10

Reeves, D. L., Winter, K. P., Bleiberg, J., and Kane, R. L. (2007). ANAM® Genogram: Historical perspectives, description, and current endeavors. *Arch. Clin. Neuropsychol.* 22, S15–S37. doi: 10.1016/j.acn.2006.10.013

Reise, S. P., and Waller, N. G. (2009). Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* 5, 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Rhoads, T., Resch, Z. J., Ovsiew, G. P., White, D. J., Abramson, D. A., and Soble, J. R. (2021). Every second counts: a comparison of four dot counting test scoring procedures for detecting invalid neuropsychological test performance. *Psychol. Assess.* 33, 133–141. doi: 10.1037/pas0000970

Rickards, T. A., Cranston, C. C., Touradji, P., and Bechtold, K. T. (2018). Embedded performance validity testing in neuropsychological assessment: potential clinical tools. *Appl. Neuropsychol.*25, 219–230. doi: 10.1080/23279095.2017.1278602

Robinson, A., Calamia, M., Penner, N., Assaf, N., Razvi, P., Roth, R. M., et al. (2023). Two times the charm: Repeat administration of the CPT-II improves its classification accuracy as a performance validity index. *J. Psychopathol. Behav. Assess.* 45, 591–611. doi: 10.1007/s10862-023-10055-7

Rodriguez, V. J., Finley, J. C. A., Liu, Q., Alfonso, D., Basurto, K. S., Oh, A., et al. (2024). Empirically derived symptom profiles in adults with attention-deficit/hyperactivity disorder: An unsupervised machine learning approach. *Appl. Neuropsychol.* 23, 1–10. doi: 10.1080/23279095.2024.2343022

Roebuck-Spencer, T. M., Vincent, A. S., Gilliland, K., Johnson, D. R., and Cooper, D. B. (2013). Initial clinical validation of an embedded performance validity measure within the automated neuropsychological metrics (ANAM). *Arch. Clin. Neuropsychol.* 28, 700–710. doi: 10.1093/arclin/act055

Roor, J. J., Peters, M. J., Dandachi-FitzGerald, B., and Ponds, R. W. (2024). Performance validity test failure in the clinical population: A systematic review and meta-analysis of prevalence rates. *Neuropsychol. Rev.* 34, 299–319. doi: 10.1007/s11065-023-09582-7

Rose, F. E., Hall, S., and Szalda-Petree, A. D. (1995). Portland digit recognition test-computerized: measuring response latency improves the detection of malingering. *Clin. Neuropsychol.* 9, 124–134. doi: 10.1080/13854049508401594

Schatz, P., and Glatts, C. (2013). "Sandbagging" baseline test performance on ImPACT, without detection, is more difficult than it appears. *Arch. Clin. Neuropsychol.* 28, 236–244. doi: 10.1093/arclin/act009

Schroeder, R. W., Martin, P. K., Heinrichs, R. J., and Baade, L. E. (2019). Research methods in performance validity testing studies: Criterion grouping approach impacts study outcomes. *Clin. Neuropsychol.* 33, 466–477. doi: 10.1080/13854046.2018.1484517

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., and Marshall, P. S. (2012). Reliable digit span: A systematic review and cross-validation study. *Assessment* 19, 21–30. doi: 10.1177/1073191111428764

Scimeca, L. M., Holbrook, L., Rhoads, T., Cerny, B. M., Jennette, K. J., Resch, Z. J., et al. (2021). Examining Conners continuous performance test-3 (CPT-3) embedded performance validity indicators in an adult clinical sample referred for ADHD evaluation. *Dev. Neuropsychol.* 46, 347–359. doi: 10.1080/87565641.2021.1951270

Scott, J. C., Moore, T. M., Roalf, D. R., Satterthwaite, T. D., Wolf, D. H., Port, A. M., et al. (2023). Development and application of novel performance validity metrics for computerized neurocognitive batteries. *J. Int. Neuropsychol. Soc.* 29, 789–797. doi: 10.1017/S1355617722000893

Sharland, M. J., Waring, S. C., Johnson, B. P., Taran, A. M., Rusin, T. A., Pattock, A. M., et al. (2018). Further examination of embedded performance validity indicators for the Conners' Continuous Performance Test and Brief Test of Attention in a large outpatient clinical sample. *Clin. Neuropsychol.* 32, 98–108. doi: 10.1080/13854046.2017.1332240

Sherman, E. M., Slick, D. J., and Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Arch. Clin. Neuropsychol.* 35, 735–764. doi: 10.1093/arclin/acaa019

Shura, R. D., Miskey, H. M., Rowland, J. A., Yoash-Gantz, R. E., and Denning, J. H. (2016). Embedded performance validity measures with postdeployment veterans: Cross-validation and efficiency with multiple measures. *Appl. Neuropsychol.* 23, 94–104. doi: 10.1080/23279095.2015.1014556

Siedlik, J. A., Siscos, S., Evans, K., Rolf, A., Gallagher, P., Seeley, J., et al. (2015). Computerized neurocognitive assessments and detection of the malingering athlete. *J. Sports Med. Phys. Fitness* 56, 1086–1091.

Singh, S., and Germine, L. (2021). Technology meets tradition: A hybrid model for implementing digital tools in neuropsychology. *Int. Rev. Psychiat.* 33, 382–393. doi: 10.1080/09540261.2020.1835839

Slick, D. J., Hoop, G., and Strauss, E. (1995). *The Victoria Symptom Validity Test.* Odessa, FL: Psychological Assessment Resources. doi: 10.1037/t27242-000

Soble, J. R., Alverson, W. A., Phillips, J. I., Critchfield, E. A., Fullen, C., O'Rourke, J. J. F., et al. (2020). Strength in numbers or quality over quantity? Examining the importance of criterion measure selection to define validity groups in performance validity test (PVT) research. *Psychol. Inj. Law* 13, 44–56. doi: 10.1007/s12207-019-09370-w

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., and Boone, K. B. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *Clin. Neuropsychol.* 35, 1053–1106. doi: 10.1080/13854046.2021.1896036

Winter, D., and Braw, Y. (2022). Validating embedded validity indicators of feigned ADHD-associated cognitive impairment using the MOXO-d-CPT. *J. Atten. Disord.* 26, 1907–1913. doi: 10.1177/10870547221112947

Zygouris, S., and Tsolaki, M. (2015). Computerized cognitive testing for older adults: a review. *Am. J. Alzheimer's Dis. Other Dement.* 30, 13–28. doi: 10.1177/1533317514522852

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 10 September 2024
DOI 10.3389/fpsyg.2024.1438830

OPEN ACCESS

EDITED BY
Jose D. Perezgonzalez,
Massey University Business School,
New Zealand

REVIEWED BY
Nicolò Zarotti,
Manchester Centre for Clinical
Neurosciences, United Kingdom
Ottavia Maddaluno,
Santa Lucia Foundation (IRCCS), Italy

*CORRESPONDENCE
Maria Grazia Vaccaro
✉ mg.vaccaro@unicz.it

†These authors have contributed equally to this work

RECEIVED 26 May 2024
ACCEPTED 13 August 2024
PUBLISHED 10 September 2024

CITATION
Maiuolo ML, Giorgini R, Vaccaro MG, Facchin A, Quattrone A and Quattrone A (2024) Assessments scales for the evaluation of health-related quality of life in Parkinson's disease, progressive supranuclear palsy, and multiple system atrophy: a systematic review. *Front. Psychol.* 15:1438830. doi: 10.3389/fpsyg.2024.1438830

Assessments scales for the evaluation of health-related quality of life in Parkinson's disease, progressive supranuclear palsy, and multiple system atrophy: a systematic review

Maria Lucia Maiuolo[1†], Roberto Giorgini[1†], Maria Grazia Vaccaro[1*], Alessio Facchin[1], Andrea Quattrone[1] and Aldo Quattrone[2]

[1]Department of Medical and Surgical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy, [2]Neuroscience Research Centre, Magna Graecia University of Catanzaro, Catanzaro, Italy

Background: The concept of wellbeing is expansive and intricate, making it challenging to define precisely. Similarly, the instruments employed to assess wellbeing are complex and multifaceted. Therefore, it is more appropriate to refer to the notion of wellbeing as Health-Related Quality of Life (HRQoL), which is the central focus of many measures used to assess the feeling of wellbeing. This review aimed to identify the tools most commonly used to evaluate HRQoL in individuals with Parkinsonism—a group of movement disorders that negatively impact the quality of life due to the intricate interplay of symptoms, socio-demographic characteristics, and psychological factors. The main aim was to assess the psychometric properties of these measures in terms of validity and reliability.

Methods: A literature review was conducted, focusing on research related to the assessment of HRQoL in connection to symptoms of Parkinsonism. This review included all studies that examined HRQoL using evaluation scales, exams, or self-reported questionnaires. The literature review was conducted using the databases Scopus and Web of Science and the search engine PubMed to identify studies published between 1996 and 2023. Only records that assessed HRQoL in individuals with Parkinson's disease and Parkinsonism were selected for evaluation.

Results: A total of 393 records were examined, and eight tools were identified as the most frequently used in the evaluation of HRQoL.

Discussion: The results show a significant gap in knowledge regarding the latent structure and measurement invariance of HRQoL measurements, which may have a significant influence on the interpretation of test outcomes. Moreover, there is a lack of clear divergent validity between HRQoL assessments and other tests used as predictors of HRQoL. This could represent a significant limitation, affecting the construct and criterion validity of HRQoL measures.

KEYWORDS

assessment tool, quality of life, systematic review, wellbeing, Parkinson's disease, progressive supranuclear palsy, multiple system atrophy, psychometrics

1 Introduction

Parkinsonism refers to a group of neurodegenerative disorders characterized by core mobility impairments resulting from pathological degeneration in specific brain regions. The most common form, Parkinson's disease (PD), affects 0.3% of the general population, with prevalence rates ranging from 1% to 5% in those aged 65 to 69 years and 1% to 3% in those aged 80 to 90 years (Arboleda-Montealegre et al., 2021; Simpson et al., 2021). The onset of symptoms in PD is associated with the loss of neurons in the nigrostriatal pathway, primarily due to an abnormal accumulation of Lewy bodies, which are complex agglomerates of proteins. The most prevalent motor symptoms (MS) include tremors, bradykinesia, stiffness, postural instability, musculoskeletal issues, gait impairment, motor fluctuations, and dyskinesia. These symptoms often lead to subsequent difficulties, such as an increased risk of falls (Kim et al., 2018; Josiah et al., 2012; Hechtner et al., 2014).

PD and other forms of Parkinsonisms are also characterized by non-motor symptoms (NMS), which include neuropsychiatric, sensory, autonomic, and sleep disorders. NMS can drastically impact patients' daily lives. For example, impairment of cognitive functioning and sensory perception can influence food consumption, leading to a lack of energy and weight loss (Akbar et al., 2015). Psychiatric comorbidities, such as anxiety and depression, are common in PD patients and worsen the prognosis, negatively affecting many aspects of their lives (D'Iorio et al., 2017; Chuquilín-Arista et al., 2021). MS and NMS vary in severity and form, depending on the specific type of Parkinsonism, and are generally associated with the partial absence of dopamine in extrapyramidal networks and reductions in white and gray matter in cortical and subcortical regions (Winter et al., 2011a,b).

Progressive supranuclear palsy (PSP) is another form of Parkinsonism, sharing symptoms with PD but also presenting distinct signs such as supranuclear vertical gaze palsy, which aids in differential diagnosis (Boxer et al., 2017, cited in Li et al., 2023). PSP is unresponsive to levodopa, resulting in a generally worse prognosis than PD due to the lack of effective therapy (Li et al., 2023). Individuals with PSP often experience substantial visual impairments combined with balance symptoms due to supranuclear center degeneration and cerebellar atrophy, respectively, which severely impact daily living and psychosocial functioning (Schrag et al., 2003). Comorbidities in PSP are related to the duration of the disease and include a broad spectrum of neuropsychiatric disorders (Schrag et al., 2006; Winter et al., 2010a,b).

Multiple system atrophy (MSA) is another rare and progressive neurodegenerative disorder within the Parkinsonism spectrum, characterized by autonomic dysfunction, cerebellar ataxia, and pyramidal symptoms (Schrag et al., 2006). MSA symptoms are linked to cortical and subcortical degeneration caused by the abnormal accumulation of Lewy bodies in nerve cells. Similar to PSP, MSA has a poor prognosis, with neuropsychiatric comorbidities often emerging as the disease progresses (Xiao et al., 2022). Individuals with MSA experience reduced psychosocial functioning due to impaired motor function and cognitive decline (Jecmenica-Lukic et al., 2018; Winter et al., 2010a,b; Du et al., 2018).

While life expectancy for individuals with PD is very close to that of the general population, MSA and PSP progress more rapidly. Currently, no therapies exist for Parkinsonism that can inhibit neurodegenerative processes, leading to the inevitable deterioration of function over time. The progression of MS and NMS often results in significant psychological consequences, which can compromise activities of daily living (ADLs), such as eating, cleaning, dressing, and working. The complex nature of network degradation associated with Parkinsonism can lead to psychiatric symptoms independent of MS and NMS severity (Simpson et al., 2021).

Parkinsonism presents a heterogeneous clinical manifestation, with symptoms that interact and collectively impact the quality of life (QoL) of those affected. QoL is a broad concept that often intersects with terms such as wellbeing and wellness. Due to this overlap, QoL is considered an umbrella term that encompasses both wellbeing and wellness (Benjamin and Looby, 1998). QoL includes several dimensions, such as spirituality, economic position, employment, interpersonal connections, and health (Benjamin and Looby, 1998). Particularly, researchers refer to "health-related quality of life" (HRQoL) to describe the ways in which perceptions of or direct repercussions from health status affect QoL (Lee et al., 2015; Global Parkinson's Disease Survey Steering Committee, 2002; Jenkins et al., 1990).

Numerous studies have documented how the complications of Parkinsonism impair HRQoL. Given the disparate clinical manifestations, scholars employ various designs to accurately measure impairment. A common approach involves quantifying symptom severity (e.g., through scale administration) as an independent variable to predict HRQoL measures. There is a well-established negative correlation between the severity of MS and HRQoL, as motor impairment directly affects ADLs and increases the risk of secondary injuries (Kim et al., 2018; Josiah et al., 2012; Hechtner et al., 2014).

Moreover, NMS appears to impact HRQoL negatively; for example, cognitive impairment, assessed through neuropsychological tests, is a significant predictor of HRQoL, **with** attentional deficits and decrease**d** executive functions lead**ing** to lower HRQoL (Leroi et al., 2011; Guo et al., 2015; Ou et al., 2017).

However, the impact of certain NMS, such as depression, anxiety, and autonomic symptoms, on HRQoL remains unclear, and whether these NMS can predict HRQoL is still questionable (Kadastik-Eerme et al., 2015; Schrag et al., 2006; Winter et al., 2010a,b; Kovács et al., 2016; Sanchez-Luengos et al., 2022; Bugalho et al., 2021; Gan et al., 2014; Li et al., 2010). One plausible reason for this discrepancy may lie in the operationalization of HRQoL, which involves the collection of techniques used to translate the construct of HRQoL and its subdomains into measurable variables. Consequently, it is crucial to analyze the validity and reliability of the measures used to estimate HRQoL in Parkinsonisms.

The construction of a tool for evaluating HRQoL in the Parkinsonism population frequently involves methods similar to those used in the validation of psychological tests. HRQoL assessments are typically self-reported and can be

classified into generic and specific types, depending on whether the items address common factors impacting HRQoL or distinctive symptoms linked with a particular condition. This study aims to clarify the psychometric characteristics of the most frequently employed tests to measure HRQoL in PD, PSP, and MSA populations. While Lewy Body Dementia, Cortico-Basal Degeneration, and Frontotemporal Dementia are additional types of Parkinsonisms, as indicated by the study's findings, these conditions are not addressed (see the Section 3).

TABLE 1 PICO.

Population of interest	Patients with PD and Parkinsonism
Intervention of interest	Use of measurement's instruments to evaluate Health-Related Quality of Life in PD and Parkinsonism
Comparison interventions	Not applicable
Outcomes	HRQoL in Parkinson's disease and Parkinsonism
Time	From 1996 to 2023
Other considerations	Sample size <50 and no power analysis carried out

2 Method

2.1 Search strategy

The search was conducted using *PubMed*, *Scopus,* and *Web of Science* with a structured search strategy, focusing on scientific articles published from 1996 to 2023. The search was completed on 20 February 2023. The following search terms were used: ((parkinson) OR (parkinsonism) AND (english[Filter])) and (((“health-related quality of life”) AND ((questionnaire) OR (“self-report”) OR (scale))) AND (english[Filter])) on *Scopus* followed by an overall search query limited to articles in English. A similar search strategy was then applied in PubMed and Web of Science using the terms ((parkinson) OR (parkinsonism)) and (((“health-related quality of life”) AND ((questionnaire) OR (“self-report”) OR (scale))).

2.2 Inclusion and exclusion criteria

The selection of research involving patients with Parkinson's disease (PD), progressive supranuclear palsy (PSP), and multiple system atrophy (MSA) for evaluating their HRQoL was conducted using the PICO (Patient, Intervention, Control, Outcome) criteria (Higgins and Green, 2011; Table 1). Studies were conducted if they lacked a power analysis or had a sample size of fewer than 50 participants. Additionally, records that did not align with the objectives of the systematic review were excluded. Due to the lack of HRQoL data on other types of parkinsonism within the research approach used in this systematic review, only PD, PSP, and MSA were considered in this analysis.

2.3 Investigated psychometric properties

Construct validity was assessed by examining the latent structure of the tests, analyzing the number of latent variables, and determining how the observed variables were associated with specific or general factors to gauge the level of agreement between the recorded findings. All articles that investigated the underlying structure of HRQoL measures were included, regardless of the specific statistical analysis methods used, such as confirmatory factor analysis (CFA) or exploratory factor analysis (EFA).

To evaluate the robustness of tests' latent structure, the literature review included all articles that analyzed measurement invariance (ME/I; Gregorich, 2006). In order to guarantee that test results remain consistent across different groups, testing ME/I is essential (Gregorich, 2006). For example, several studies have reported differences in HRQoL between men and women (Ophey et al., 2018). Another reason for including ME/I studies is that many of the measurements have undergone cross-cultural validation. This validation is necessary to guarantee that a test is appropriately adapted to different cultural contexts, making ME/I studies crucial.

While a clear and replicable latent structure is necessary, it is not sufficient to fully establish the construct validity of a test. Therefore, all paradigms of convergent and divergent validity for HRQoL measure were included, with particular attention given to divergent validity due to the theoretical overlap of constructs related to QoL. The Cronbach's alpha and other reliability indices of each test were also reported. When considering the usage of Cronbach's alpha, it is important to focus on the assumption of tau-equivalence in the latent structure. Significant emphasis was placed on this assumption check, as it is essential for ensuring unbiased results when using Cronbach's alpha (Flake et al., 2017).

The criterion validity of HRQoL tests was reported by assessing the main predictors of HRQoL, as well as by measuring comorbidities (e.g., depression) and MS and NMS scales. This review specifically focused on the rationale behind these predictions rather than investigating the predictor measures' psychometric properties. Additionally, validation techniques that followed the Item Response Theory framework were also included in the analysis.

3 Results

3.1 Identification of records

The PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) diagram (Page et al., 2021a,b) summarizes the search results, article screening process, and exclusions. A total of 1,159 records were identified from the search engines and databases, comprising 85 from Scopus, 573 from PubMed, and 501 from Web of Science (Figure 1). After the initial screening, 572 records and 277 duplicates were excluded (Figure 1). Further exclusions were made for studies focused on caregivers ($n = 43$),

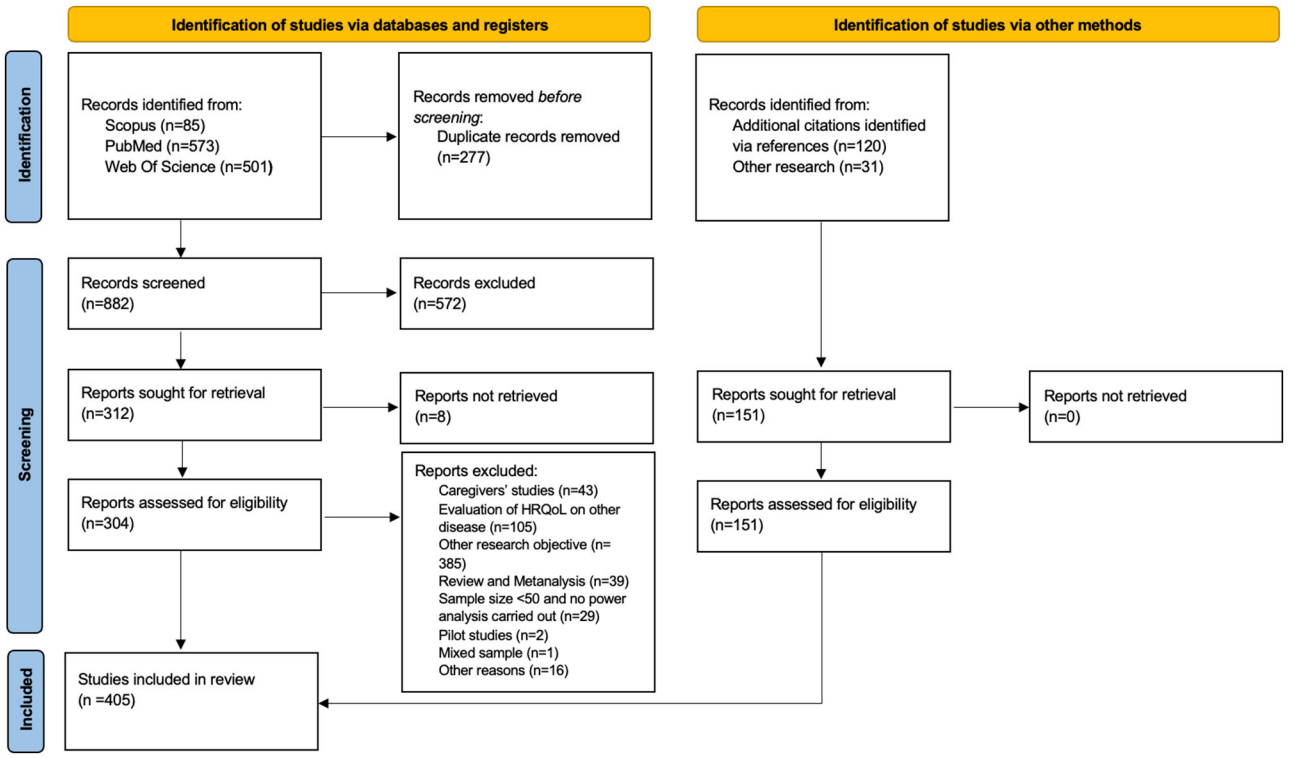


FIGURE 1
PRISMA flow diagram. PRISMA diagram for the selection of studies.

HRQoL in other diseases ($n = 105$), irrelevant research aims ($n = 385$), and systematic reviews ($n = 39$).

During the second screening process, additional exclusions were made for non-retrievable studies ($n = 8$), studies not aligned with our research objective or identified as pilot studies ($n = 2$), studies with mixed samples ($n = 1$), and those with a small sample size and lacking a clearly reported power analysis ($n = 29$). Moreover, 16 studies were excluded due to the complete absence of a method description or unclear methodology. An additional 151 records were identified through references or other research efforts. Finally, a total of 405 studies were included in this systematic review.

3.2 Identification of measures

The total number of HRQoL assessment tools and other tests applied in specific study designs was $n = 121$ (see Table 2).

The main instruments found in more than 3 records include 25 tools, with the most commonly used being:

- Specific instruments for evaluating HRQoL in Parkinsonisms: PDQ-39, PDQ-8, PDQL, and SCOPA-PS.
- Generic instruments for evaluating HRQoL: SF-36, SF-12, EuroQol-5 (EQ-5D-5L and EQ-5D-3L), and NHP.
- Instruments for evaluating predictors of HRQoL outcomes in PD, PSP, and MSA: MDS-UPDRS, CISI-PD, H&Y Stage, S&E, UMSARS, NMS-Quest, and NMSS.

These instruments are illustrated in Figure 2.

3.3 Specific instruments to evaluate HRQoL in PD, PSP, and MSA

3.3.1 PDQ-39

The 39-item Parkinson's Disease Questionnaire (PDQ-39) is one of the most widely used instruments for evaluating HRQoL in individuals with Parkinson's disease (Peto et al., 1995). It comprises 39 items that are divided into eight scales: mobility (eight items), Activities of Daily Living (ADL; six items), emotional wellbeing (six items), stigma (four items), social support (three items), cognitions (four items), communication (three items), and bodily discomfort (three items). Each item is scored on a scale from 0 to 4, and the total score, which is called the PDQ-39 Summary Index (PDQ-39SI), ranges from 0 to 100. The PDQ-39 is available in 14 languages: English, Spanish, American, Greek, Chinese, Singaporean, Ecuadorian, French, Brazilian, Iranian, Portuguese, Korean, Estonian, and Italian (Galeoto et al., 2018). It is also used in populations with MSA and PSP (Schrag et al., 2006).

Several studies have assessed the reliability of the PDQ-39, with generally acceptable values reported. However, most applications of Cronbach's alpha did not account for tau-equivalence, which

10.3389/fpsyg.2024.1438830

TABLE 2 List of the instruments found for measuring HRQoL and predictors.

	Physical	Mental	Social	ADL	Overall	Other	References
15D generic instruments	✓	✓	✓	✓	-	-	Sintonen, xbib1994
Activities-Specific Balance Confidence Scale	-	-		✓	-	-	Powell and Myers, 1995
ADL (Activities of Daily Living)	-	-	-	✓	-	-	Lawton and Brody, 1969
AIS (Athens Insomnia Scale)	✓	-	-	-	-	-	Soldatos et al., 2000
Apathy Scales	-	✓	-	-	-	-	Starkstein et al., 1992
BAI (Beck Anxiety Inventory)	-	-	-	-	-	✓	Beck et al., 1993
Barthel Index of ADL	-	-	-	✓	-	-	Mahoney and Barthel, 1965
BBS (Berg Balance Scale)	-	-	-	-	-	✓	Berg et al., 1995
BDI (Beck Depression Inventory)	-	-	-	-	-	✓	Beck et al., 1987
Behave-AD (Behavior Pathology in Alzheimer's Disease Rating Scale)	-	-	-	-	-	✓	Reisberg et al., 1987
BELA-P-k (Belastungsfragebogen Parkinson's kurzversion)	✓	✓	✓	-	-	-	Ringendahl et al., 2000
BFAS (Big Five Aspects Scale)	-	-	-	-	-	✓	DeYoung et al., 2007
BIS-11 (Barratt Impulsiveness Scale)	-	-	-	-	-	✓	Patton et al., 1995
CESD (Center for Epidemiologic Studies Depression Scale)	-	-	-	-	-	✓	Radloff, 1977
CGI (Clinical Global Impression of Change)	-	-	-	-	-	✓	Weitkunat et al., 1993
CISI-PD (Clinical Impression Of Severity Index-Parkinson's Disease)	✓	✓	-	✓	-	✓	Martínez-Martín et al., 2003
COMPASS (Composite Autonomic symptom scale)	-	-	-	-	-	✓	Suarez et al., 1999
Composite International Diagnostic Interview Short Form for Major Depression	-	-	-	-	-	✓	Kessler et al., 1998
CISS (Coping Inventory for Stressful Situations)	-	-	-	-	-	✓	Endler and Parker, 1999
CSQ (Coping Strategies Questionnaire)	-	-	-	-	-	✓	Rosenstiel and Keefe, 1983
DASS-21 (Depression Anxiety Stress Scales)	-	-	-	-	-	✓	Lovibond and Lovibond, 1995
ESES (Exercise Self-Efficacy Scale)	-	-	-	-	-	✓	Kroll et al., 2007
ESS (Epworth Sleepness Scale)	-	-	-	-	-	✓	Johns, 1991
PDCS (European Parkinson's Disease Association Sponsored)	-	-	✓	✓	-	✓	Stocchi et al., 2018
EUROQoL	✓	✓	✓	✓	✓	-	Euroqol Group, 1990
FACIT (Functional Assessment of Chronic Illness Therapy)	✓	✓	✓	✓	✓	✓	Webster et al., 1999
FBI (The 24-item Frontal Behavioral Inventory)	-	-	-	-	-	✓	Kertesz et al., 1997

(Continued)

TABLE 2 (Continued)

	Physical	Mental	Social	ADL	Overall	Other	References
FES (Falls Efficacy Scale)	-	✓	-	✓	-	-	Tinetti et al., 1990
FFRT (Forward Functional Reaching Test)	-	-	-	-	-	✓	Duncan et al., 2014
FKV-LIS-SE (the Freiburg Coping with Disease Questionnaire)	-	-	-	-	-	✓	Muthny, 1989
FOG-Q (Freezing of Gait Questionnaire)	✓	-	-	-	✓	-	Giladi et al., 2000
FSI (Fatigue Severity Inventory)	-	-	-	-	-	✓	Lee et al., 1991
FSQ (Functional Status Questionnaire)	-	-	-	-	-	✓	Jette et al., 1986
FSS (Fatigue Severity Scale)	✓	-	-	✓	-	-	Krupp et al., 1989
GDS-15 (Geriatric Depression Scale)	-	-	-	-	-	✓	Yesavage et al., 1982
GSE (General Self-Efficacy Scale)	-	-	-	-	-	✓	Schwarzer and Jerusalem, 1999
German Essen Coping Questionnaire	-	-	-	-	-	✓	Franke et al., 2000
HandY stage (Hoehn and Yahr stage)	-	-	-	-	-	✓	Hoehn and Yahr, 1967
HADS (Hospital Anxiety and Depression Scale)	-	-	-	-	-	✓	Zigmond and Snaith, 1983
HAM-A (Hamilton Anxiety Rating Scale)	-	-	-	-	-	✓	Hamilton, 1959
HAM-D (Hamilton Depression Rating Scale)	-	-	-	-	-	✓	Hamilton, 1960
HUI-3 (Health Utilities Index Mark)	-	-	-	-	✓	-	Furlong et al., 1998
HWS (Holistic Wellbeing Scale)	✓	✓	-	-	✓	✓	Chan et al., 2014
IADL (Instrumental Activities of Daily Living)	-	-	-	✓	-	-	Lawton and Brody, 1969
Impulsive-Compulsive Disorders in Parkinson's Disease	-	-	-	-	-	✓	Weintraub et al., 2009
IPAQ (International Physical activity Questionnaire)	-	-	-	✓	-	✓	Cardol et al., 2001
IQCODE ('e Informant Questionnaire on Cognitive Decline in the Elderly)	-	-	-	-	-	✓	Cherbuin and Francis Jorm, 2010
King's Parkinson's Disease Pain Scale	✓	-	-	-	-	✓	Chaudhuri et al., 2015
LARS (Lille Apathy Rating Scale)	-	✓	-	-	-	-	Sockeel et al., 2006
Leed Anxiety and Depression scale	-	-	-	-	-	✓	Snaith et al., 1976
Livingston's Insomnia Scale	-	-	-	-	-	✓	Livingston et al., 1993
LOT-R (Life Orientation Test Revised)	-	-	-	-	-	✓	Scheier and Carver, 1985
MAAS (Mindful Attention Awareness Scale)	-	-	-	-	-	✓	Brown and Ryan, 2003

(Continued)

TABLE 2 (Continued)

	Physical	Mental	Social	ADL	Overall	Other	References
MADRS (Montgomery and Asberg Depression Rating Scale)	-	-	-	-	-	✓	Neumann and Schulte, 1989
MSS (Marital Satisfaction Scale)	-	-	-	-	-	✓	Roach et al., 1981
MDRS (Modified Dyskinesia Rating Scale)	✓	-	-	-	-	-	Goetz et al., 1994
MDS-UPDRS (Movement Disorder Society- Unified Parkinson's Disease Rating Scale)	-	-	-	-	-	✓	Goetz et al., 2008
Mini-BESTest (The Mini-Balance Evaluation Systems Test)	–	-	-	-	-	✓	Horak et al., 2009
MAS-QoL(MSA health-related Quality of Life scale)	✓	✓	✓	-	-	-	Schrag et al., 2007
MSPQ (Modified Somatic Perception Questionnaire)	-	-	-	-	-	✓	Main, 1983
NAS (Nottingham Adjustment Scale)	-	-	-	-	-	✓	Dodds et al., 1991
NEURO-QOL (Quality of Life in Neurological Disorders)	✓	✓	✓	✓	✓	-	Gershon et al., 2012
NHP (Nottingham Health Profile)	✓	✓	✓	✓	✓	-	Hunt et al., 1985
NMS-Quest (Non-Motor Symptoms Questionnaire)	✓	✓	-	-	-	✓	Romenets et al., 2012
NMSS (Non-Motor Symptom Scale)	-	✓	-	-	-	✓	Chaudhuri et al., 2007
OARS (The Older Americans Resources and Services)	✓	✓	✓	✓	✓	-	Fillenbaum and Smyer, 1981
PAS (Parkinson's Anxiety Scale)	-	-	-	-	-	✓	Leentjens et al., 2014
PANAS (The Positive and Negative Affect Schedule)	-	✓	-	-	-	-	Watson et al., 1988
PCIG (Patient Global Impression of Change)	-	-	-	-	-	✓	Ferguson and Scheman, 2009
PCQ-PD (Patient-Centered Questionnaire for PD)	-	-	-	-	-	✓	van der Eijk et al., 2012
PDQ-39 (Parkinson's Disease Questionnaire-39 item)	✓	✓	✓	✓	✓	✓	Peto et al., 1995
PDQ-8 (Parkinson's Disease Questionnaire-8 item)	✓	✓	✓	✓	✓	✓	Jenkinson et al., 1997
PDQL (Parkinson's disease quality of life questionnaire)	✓	✓	✓	-	✓	✓	de Boer et al., 1996
PDQualif (the Parkinson's Disease Quality of Life Scale)	-	✓	✓	✓	-	✓	Welsh et al., 2003
PDSS (Parkinson's disease sleep scale)	-	-	-	-	-	✓	Chaudhuri et al., 2002
PWI-A (Personal Wellbeing Index-Adult)	✓	✓	✓	✓	✓	✓	Lau et al., 2005
PFS-16 (Parkinson's Fatigue Scale)	-	-	-	-	-	✓	Brown et al., 2005

(Continued)

TABLE 2 (Continued)

	Physical	Mental	Social	ADL	Overall	Other	References
PGIC (Additional secondary measures of Patient Global Impression of Change)	✓	✓	-	-	✓	-	Hurst and Bolton, 2004
PHQ-9 (Patient Health Questionnaire)	-	-	-	-	-	✓	Kroenke et al., 2001
PILL Questionnaire (Impact of Cognitive Dysfunction on Daily Living Activities)	-	✓	-	✓	-	-	Dubois et al., 2007
PROMIS (The patient-reported outcomes measurement information system)	✓	✓	✓	-	✓	-	Ader, 2007
PSP-QoL (Progressive Supranuclear Palsy Rating Scale)	✓	✓	✓	✓	-	-	Schrag et al., 2006
PSP-RS (Progressive Supranuclear Palsy Rating Scale)	✓	-	-	-	-	✓	Golbe and Ohman-Strickland, 2007
PSQI (Pittsburgh Sleep Quality Index)	-	-	-	✓	-	✓	Buysse et al., 1989
PWS (Psychological Wellbeing Scale)	-	-	-	-	✓	✓	Ryff, 1989
QOL-AD (QOL Alzheimer's Disease)	-	-	-	✓	-	✓	Logsdon et al., 1999
QUEST (Quality of Life in Essential Tremor Questionnaire)	✓	✓	✓	✓	✓	✓	Tröster et al., 2005
RAD (Rapid Assessment of Disability Scale)	-	-	✓	✓	✓	-	Martinez-Martin et al., 2005
RBDSQ (REM Sleep Behaviour Disorder Symptoms Questionnaire)	-	-	-	-	-	✓	Stiasny-Kolster et al., 2007
RCSQ (Richards–Campbell Sleep Questionnaire)	-	-	-	-	-	✓	Richards, 1987
RSE (Rosenberg Self-Esteem Scale)	-	-	-	-	-	✓	Rosenberg, 1965
Ryff's scale of Psychological Wellbeing	-	-	-	-	✓	✓	Ryff, 1989; Ryff and Keyes, 1995
SandE (Schwab and England scale)	-	-	-	✓	-	-	Schwab and England, 1969
SAMS (German Stendal Adherence with Medication Score)	-	-	-	-	-	✓	Franke and Jagla-Franke, 2020
SCOPA-PS (Scale for Outcomes in Parkinson's Disease -Psychosocial questionnaire)	-	✓	✓	-	-	-	Marinus et al., 2003
SDS (Self-rating Depression Scale)	-	-	-	-	-	✓	Zung, 1965; Biggs et al., 1978
SEE (Self-Efficacy for Exercise scale)	-	-	-	-	-	✓	Resnick and Jenkins, 2000
SCOPA-AUT (Self-reported Autonomic Symptoms in Parkinson's Disease)	-	-	-	-	-	✓	Visser et al., 2004
SF-12 (Short-Form Health Survey 12 item)	✓	✓	✓	✓	✓	-	Ware et al., 1996

(Continued)

TABLE 2 (Continued)

	Physical	Mental	Social	ADL	Overall	Other	References
SF-36 (Short-Form Health Survey 36 item)	✓	✓	✓	✓	✓	-	Ware and Sherbourne, 1992
SF-6D (Short-Form Six Dimension)	✓	✓	✓	✓	✓	-	Brazier et al., 2002
Short Social Support Questionnaire	-	-	✓	-	-	-	Jahanshahi and Marsden, 1988; Sarason et al., 1983
SIP (Sickness Impact Profile)	✓	-	✓	✓	-	-	Bergner et al., 1976
SIPA (Social influences on physical Activity questionnaire)	-	-	-	-	-	✓	Chogahara, 1999
SOC-29 (Sense of Coherence Scale)	-	-	-	-	-	✓	Antonovsky, 1972
SOFAS (Social and occupational functioning assessment scale)	-	-	✓	✓	-	✓	Rybarczyk, 2011
SPMSQ (The Short Portable Mental Status Questionnaire)	✓	✓	✓	✓	-	✓	Pfeiffer, 1975
STAI (State Trait Anxiety Inventory)	-	-	-	-	-	✓	Spielberger et al., 1970
UMSARS (Unified Multiple System Atrophy Rating Scale)	✓	-	-	✓	-	✓	Wenning et al., 2004
UPDRS (Unified Parkinson's Disorder Rating Scale)	-	-	–	-	-	✓	Fahn, 1987
WCQ (Ways of Coping Questionnaire)	-	-	-	-	-	✓	Folkman and Lazarus, 1985
WHO-5 (World Health Organization Well Being Index 5 item)	-	-	-	-	-	✓	World Health Organization, 1999
WHO-DAS (World Health Organization -Disability Assessment Schedule)	-	-	✓	✓	-	✓	Ustün et al., 2010
WHO-DAS II (World Health Organization -Disability Assessment Schedule-II)	✓	-	✓	✓	✓	✓	World Health Organization, 1999
WHOQOL-100 (The World Health Organization Quality of Life)	✓	✓	✓	✓	✓	✓	The WHOQOL Group, 1998; Whoqol Group, 1998
WPAI-GH (The Work Productivity and Activity Impairment Questionnaire-General Health)	-	-	-	-	-	✓	Reilly et al., 1993
AES (Apathy Evaluation Scale)	-	-	-	-	-	✓	Marin et al., 1991; Santangelo et al., 2014
ZUF-8 (Patient Satisfaction Questionnaire)	-	-	-	-	-	✓	Schmidt et al., 1989
Starkestain's Apathy Scale (Structured Clinical Interview for Apathy)	-	-	-	-	-	✓	Starkstein et al., 2001
Zung-Depression Inventory-Self Rating Depression Scale	-	-	-	-	-	✓	Zung, 1965, 1972

could affect the accuracy of the reliability estimates. The construct validity of the PDQ-39 appears to be somewhat clear. Some reports indicate that some items may load on different scales than originally intended, raising questions about the clarity of the questionnaire's structure (Schönenberg et al., 2023). Furthermore, the latent structure of the PDQ-39 has been evaluated in only

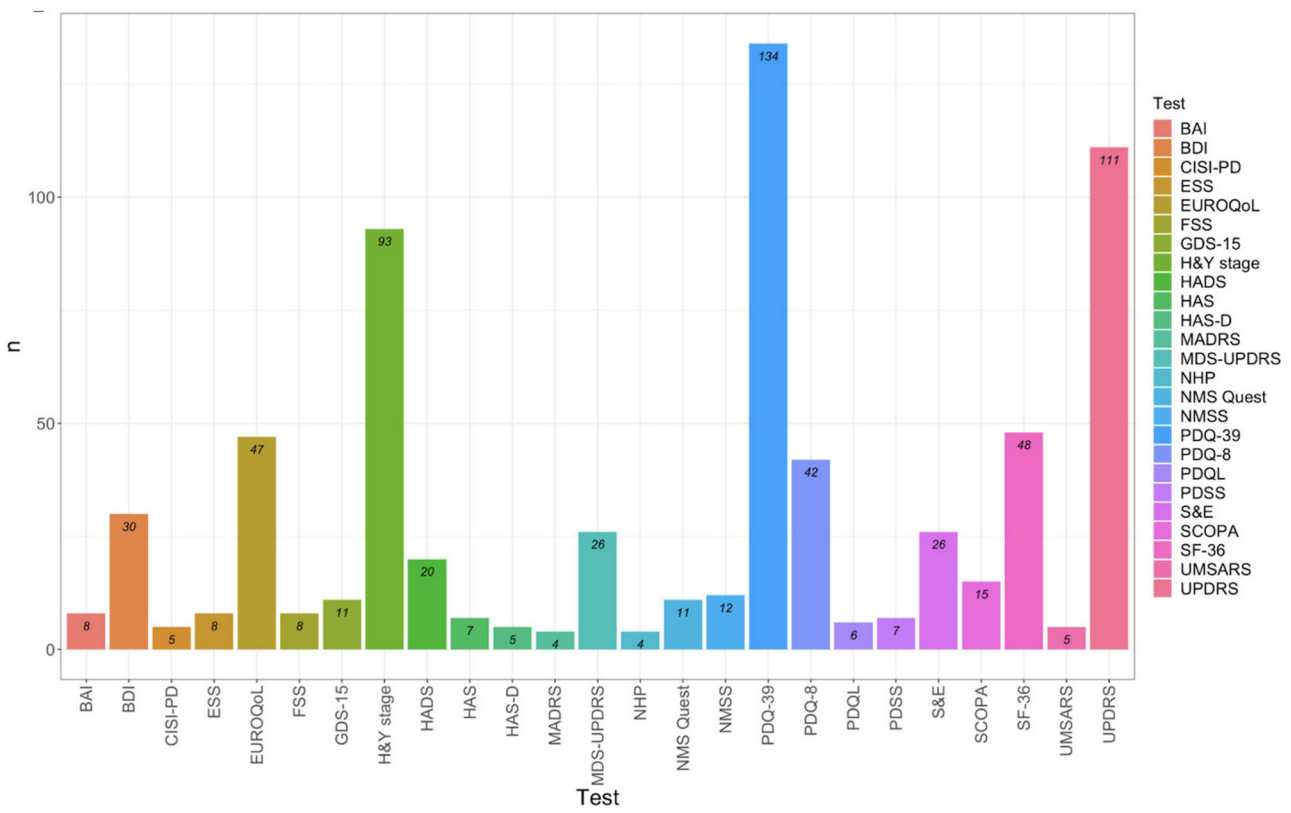


FIGURE 2
Test identified in more than 3 records. BAI, Beck Anxiety Inventory; BDI, Beck Depression Inventory; CISI-PD, Clinical Impression of Severity Index for Parkinson's Disease; ESS, Empworth Sleepiness Scale; EUROQoL, EuroQol group's test; FSS, Fatigue Severity Scale; GDS-15, Geriatric Depression Scale; HandY stage, Hoehn and Yahr stage; HADS, Hospital Anxiety and Depression Scale; HAM-A, Hamilton Anxiety rating scale; HAM-D, Hamilton Rating scale for depression; MADRS, Montgomery-Asberg Depression Rating Scale; MDS-UPDRS, Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale; NHP, Nottingham Health Profile; NMS-Quest, Non-Motor Symptoms Questionnaire; NMSS, Non-motor symptoms scale; PDQ-39, Parkinson's Disease Questionnaire 39 items; PDQ-8, Parkinson's Disease Questionnaire 8 items; PDQL, Parkinson's Disease Quality of Life Questionnaire; PDSS, Parkinson's Disease SleepScale 2nd version; SandE: Schwab and England Scale; SCOPA, Scales for Outcomes in Parkinson's Disease-Psychosocial Functioning; SF-36, 36 and 12 items Short-Form Health Survey; UMSARS, Unified Multiple System Atrophy Rating Scales; UPDRS, Unified Parkinson's Disease Rating Scale.

a few studies, which shows that the different latent variables are strongly related (Schönenberg et al., 2023). Moreover, while several studies have reported gender differences in PDQ-39 scores (e.g., Ophey et al., 2018), the ME/I of the PDQ-39 scores has not been thoroughly examined. Furthermore, no studies were found that assessed ME/I of the PDQ-39 across different countries.

3.3.2 PDQ-8

The 8-item Parkinson's Disease Questionnaire (PD1-8) consists of eight items derived from PDQ-39 using Principal Component Analysis (PCA), with item selection based on item-factor correlations (Jenkinson et al., 1997). The total score, known as PDQ-8 SI, ranges from 0 to 100 (Luo et al., 2009). The PDQ-8 has been reported to be reliable (Chen et al., 2017; Martínez-Martín et al., 2003; Alvarado-Bolaños et al., 2015; Tan et al., 2004; Li et al., 2023; Franchignoni et al., 2008), but Cronbach's alpha varies across studies, typically ranging between 0.72 and 0.8.

The rationale for creating a short form of the PDQ-39 originates from the results of the hierarchical principal component analysis conducted by Jenkinson et al. (1997). This analysis extracted one component, suggesting that the eight dimensions of the PDQ-39 reflect a high-order factor. Furthermore, these results support the interpretation of item sum scores in both the PDQ-39 and PDQ-8 as general indices of HRQoL. However, results from Rasch analysis suggest that the PDQ-8 may not effectively differentiate HRQoL as a continuum and that its structure is not unidimensional (Franchignoni et al., 2008). Similar to the PDQ-39, no records were found that checked for tau-equivalence despite widespread reporting of Cronbach's alpha. Additionally, no studies were identified that tested for ME/I.

3.3.3 PDQL

The Parkinson's Disease Quality of Life Questionnaire (PDQL) was developed by de Boer et al. (1996) and consists of 37 items divided into four dimensions: *Parkinsonian symptoms*, *systemic symptoms*, *emotional functioning*, and *social functioning*. The total score is calculated by summing the scores of each dimension, with a higher score indicating better HRQoL. The four dimensions of the PDQL were identified through EFA, and the reliability was

evaluated by testing the internal consistency of subscales using Cronbach's alpha (de Boer et al., 1996).

Moreover, the convergent validity of the PDQL was tested by examining the correlation between its subscales and the Medical Outcome Studies-24 (MOS-24), a generic wellness test. In the initial validation and subsequent studies, PDQL demonstrated good reliability, with Cronbach's alpha exceeding 0.80 for each subscale and 0.90 for the total score (de Boer et al., 1996). However, despite this high internal consistency, the correlation between the PDQL and MOS-24 subscales was weak, particularly for the "Social Functioning" dimension, where the correlation ranged from 0.13 to 0.43.

Moreover, the factor structure of the PDQL includes complex loadings, meaning that some items appear to reflect more than one dimension simultaneously. No confirmatory studies were found to validate the PDQL's structure or justify the sum score. Furthermore, cross-cultural validation studies did not include ME/I across different countries or genders.

3.3.4 SCOPA-PS

The Scales for Outcomes in Parkinson's Disease-Psychosocial (SCOPA-PS) is a self-report tool composed of 11 items on a 0–3 Likert scale (Marinus et al., 2003). The outcome of the SCOPA-PS is summarized in a summary index (SI), where a higher score indicates poorer HRQoL. The content of items reflects various social scenarios of daily living in which patients may have experienced suffering or difficulty in the previous month. Since its initial validation, SCOPA-PS has demonstrated good convergent validity through correlations with PDQ-39 and other generic tests. The factor structure appears unidimensional, although some indices, such as the RMSEA, are mediocre (RMSEA $>$ 0.08). Reliability is generally good (Soulas et al., 2016; Virués-Ortega et al., 2009). However, Virués-Ortega et al. (2009) suggest that a two-dimension structure could be a possible factor solution. Despite the good reliability, there is a lack of evidence for tau equivalence in SCOPA-PS. Regarding divergent validity, SCOPA-PS showed a high correlation with anxiety and depression scales. No studies on ME/I across gender or culture were found.

3.4 Generic instruments to evaluate HRQoL in PD, PSP, and MSA

Based on the reviewed literature, the following section describes the principal generic tools used to assess HRQoL in PD, MSA, and PSP. Specifically, this section focuses on articles that included studies for these tools in PD, PSP, and MSA. The information provided here outlines the psychometric properties of the generic tools used in these patient populations.

3.4.1 SF-36

The Short-Form Health Survey (SF-36) (Ware and Sherbourne, 1992) was designed to evaluate general HRQoL and includes eight scales that assess various health concepts, selected from 40 measured concepts by the Medical Outcome Study (MOS) through 36 items (Brown et al., 2009).

3.4.2 SF-12

The SF-12 is the shortest form of the SF-36 questionnaire, comprising 12 items that evaluate the same eight dimensions as the SF-36. These outcomes are represented only by the PCS and MCS, with higher scores on these subscales indicating better HRQoL (Ware et al., 1996). The factor structure, reliability, convergent and divergent validity of the SF-12 in a sample of PD patients were analyzed by Jakobsson et al. (2012). The results showed good reliability for the two subscales, but the CFA outcomes showed inadequate fit indices. Conversely, Hagell and Westergren (2011) demonstrated that the structure of SF-12 showed a good fit through Item Response Theory validation, although some items showed misfits. No validation studies on PSP and MSA were found. Additionally, tau-equivalent and ME/I in PD, PSP, and MSA samples have never been investigated in any studies.

3.4.3 EuroQol-5D

The EuroQol-5 (Euroqol Group, 1990) is a generic instrument used to measure the quality of life in three different ways: a descriptive system assessing health status across five dimensions, a 0–100 Visual Analogue Scale (VAS) for self-rating of one's health, and an index score reflecting the utility or preference measures. The five dimensions (5D) can be assessed using three levels (EQ-5D-3L) or five levels (EQ-5D-5L) scales or through the EQ-5D Visual Analogue Scale (VAS). The value range extends from 1.0 ("*perfect health state*") to −1.0 ("*death*"; Martínez-Martín et al., 2003). The EQ-5D is frequently used to assess HRQoL in PD (Visser et al., 2008), PSP (Picillo et al., 2019), and MSA (Winter et al., 2011a,b), but no specific validation of this scale for these populations was found.

3.4.4 NHP

The Nottingham Health Profile (NHP) is a test used to assess health status (Hunt et al., 1985). Developed in the United Kingdom, it evaluates the perception of health-related problems in physical, social, and emotional domains. The NHP is brief, easy to complete, generic, and reliable, with extensive validation of its psychometric properties across different populations (Sitzia et al., 1998; Karlsen et al., 2000). The NHP consists of two parts:

- Part 1: It contains 38 dichotomous items covering six health dimensions: *pain* (eight items), *emotional reactions* (nine items), *social isolation* (five items), *physical mobility* (eight items), *energy* (three items), and *sleep* (five items). Respondents answer "*yes*" or "*no*" to 38 questions. Each dimension is weighted, with scores ranging from 0 ("*good health*") to 100 ("*poor health*"; Savci and Sendir, 2009). It is common for NHP-1 to be utilized alone (Savci and Sendir, 2009).
- Part 2: It consists of seven statements related to areas of daily life affected by health: *paid employment*, *personal relationships*,

jobs around the house, sex life, hobbies, interests, social life, and holidays (Hunt et al., 1985).

For each dimension, the highest score is 100, while the generic highest score is 600; high NHP scores predict low HRQoL levels (Sitzia et al., 1998). No validation records for the NHP in Parkinson's and Parkinsonism populations were found, except one. Despite a small sample size, Hagell et al. (2003) suggested that the NHP requires further development analysis in the PD population due to some item misfits and low convergent validity with PDQ-39 subscales. Furthermore, Cronbach's alpha for the social isolation, energy, and sleep subscales was poor ($0.63 < \text{alpha} < 0.78$; Hagell et al., 2003), and there is no information about tau-equivalence. No ME/I studies across clinical populations and healthy subjects were found.

3.5 Instruments to evaluate HRQoL's predictors in PD, PSP, and MSA

In this review, we focus on the psychometric properties of HRQoL tests. Therefore, in this section, we do not discuss the psychometric properties of the tools listed below. The purpose of describing the predictor measurements is to clearly identify the construct they examine to analyze the criterion validity of HRQoL measurements. Only a qualitative description of the main impacts identified by tests and questionnaires in the literature is provided, as a meta-analytic approach would be more suitable for estimating quantitative information about the different effects.

3.5.1 MDS-UPDRS

The MDS-UPDRS is the most recent version of the Unified Parkinson's Disease Rating Scale (UPDRS), originally developed in the 1980s by the Movement Disorder Society. It was designed to evaluate both MS and NMS symptoms of Parkinson's disease through a combination of clinical interviews and self-reported scales. The MDS-UPDRS items range from 0 (*normal*) to 4 (*severe*) and are organized in four subscales:

- Part 1 (nmEDL): 13 items (six semi-structured and seven self-reported) evaluate non-motor experiences of daily living.
- Part 2 (mEDL): 13 self-reported items assess motor experiences of daily living.
- Part 3 (mEx): 18 items summarize motor examination.
- Part 4 (mCompl): Six items (a semi-structured interview) evaluate motor complications.

This instrument is available in multiple languages, including English, French, Hungarian, German, Estonian, Italian, Russian, Spanish, and Slovak (Skorvanek et al., 2018; Martínez-Martín et al., 2014). In summary, the MDS-UPDRS is a multidimensional hybrid battery that assesses various characteristics related to HRQoL in PD patients.

The MDS-UPDRS was used to predict HRQoL in 26 studies (Zipprich et al., 2021; Nakano et al., 2021; Ueno et al., 2020). Our research indicates that MDS-UPDRS subscales are positively related to PDQ-39 and PDQ-8 scores, while they are negatively related to generic tests, consistent with the interpretation of the scales. In the majority of records, scholars applied linear regression to quantify the influence of disease severity on HRQoL. The results of these studies suggest a general consensus on the negative impact of both objective and self-reported motor symptoms.

However, the significance of the impact of non-motor symptoms on HRQoL varied across studies. Moreover, the proportion of HRQoL variance explained also differed across studies (e.g., Ueno et al., 2020; Simpson et al., 2014; Li et al., 2010; Grimbergen et al., 2013).

3.5.2 H&Y stage scale

The Hoehn and Yahr (1998) stage is a clinical grading system based on five categories used to describe the severity of motor impairment in Parkinson's disease. The degree of motor impairment is reflected in the score, which ranges from 1 to 5. There is a consistent negative correlation between the H&Y stage and HRQoL, regardless of the statistical analysis method used. Additionally, HRQoL appears to decline progressively with increasing severity indicated by the H&Y stage, with few exceptions, such as the findings by Ophey et al. (2018) (e.g., Fereshtehnejad et al., 2014a,b; Guo et al., 2015; Hechtner et al., 2014).

3.5.3 CISI-PD

The Clinical Impression of Severity Index for Parkinson's Disease is a clinical interview that assesses four dimensions related to PD symptomatology: *motor signs*, *disability, motor complications,* and *cognitive status* (Martinez-Martin et al., 2005).

The CISI-PD includes a motor examination, with each dimension's outcomes ranging from 0 (*no improvement*) to 6 (*very severe deficit*). Additionally, the four subscales can also be summarized into a single index, with severity levels categorized as mild (1–7), moderate (8–14), and severe ($\geq$15). The CISI-PD scores negatively impact HRQoL, independent of the measure used.

For example, Norlin et al. (2023) demonstrated that PD patients classified according to CISI-PD sum scores displayed HRQoL levels (measured with PDQ-8 score and EQ-5D scores) corresponding to their severity. This effect appears to affect men and women across different age ranges but only moderately explains the variance in HRQoL. Disability and cognitive status showed higher effect sizes, but these results should be interpreted with caution due to the sum-score limitation of PDQ-8 and the absence of EQ-5D validation in PD populations. These effects appear consistent across the studies included, confirming the relationship between objective symptoms and self-reported HRQoL measures.

3.5.4 NMS-Quest and NMSS

The Non-Motor Symptoms Questionnaire (NMS-Quest) and the Non-Motor Symptom Scale (NMSS) are tools for screening and evaluating the non-motor features of PD (Chaudhuri et al., 2007). Both tests include 30 items divided into 9 domains: *cardiovascular, sleep disorders, mood/cognition, hallucinations, attention/memory, gastrointestinal, urinary, sexual dysfunction, and miscellaneous.*

The NMS-Quest includes 30 dichotomous items (*yes*/*no*) and is self-administrated, while the NMSS is a clinical interview in which the examiner rates the frequency and severity of symptoms reported by the patient (frequency ranges from 1 to 4 and severity from 0 to 3). The items can be summed in a general composite score or separated by domains for each questionnaire. High scores on the NMS-Quest and NMSS are associated with worse overall scores. However, replicability of the effects of non-motor symptoms on HRQoL was not consistently found across studies included in the reviews (e.g., Chaudhuri et al., 2013; Rosqvist et al., 2021; Shalash et al., 2018).

3.5.5 S&E scale

The Schwab and England (1969) scale evaluates the percentage of impairment in patients' independence during activities of daily living (ranging from 100% = no impairment to 0% = completely dependent and comatose). The impact of the S&E scale on HRQoL appears independent of tools used to assess it and is strongly associated with subscales that evaluate mobility and ADL (e.g., PDQ-39 mobility and ADL subscales; Schrag et al., 2006). Moreover, the S&E scale scores correlate with psychological sunctioning and emotional wellbeing (PDQ-39 subscale and EQ-5D subscale, respectively), suggesting an association between motor impairment and these domains, further supporting a nomological network (Schrag et al., 2006).

3.5.6 UMSARS

The Unified Multiple System Atrophy Rating Scales (UMSARS) is a clinical interview designed to assess the fundamental symptoms of MSA, categorizing the severity of its features on a scale from 0 (no impairment) to 4 (severe impairment; Wenning et al., 2004). The multidimensional tool includes four sections: *History of Disease* (Part I), *Motor Examination* (Part II), *Autonomic Examination* (Part III), and *Global Disability Scale* (Part IV). Only three studies have employed UMSARS as a predictor of HRQoL (Winter et al., 2011a,b; Schrag et al., 2006; Xiao et al., 2022). All these studies reached the same conclusion: there is a negative relationship between HRQoL and UMSARS scores. Specifically, UMSARS Part-II, Part-IV, and the total score appear to be strong predictors of SF-36 and EQ-5D outcomes, regardless of the type of MSA (whether cerebellar or Parkinsonian).

3.6 Comorbidities assessment tools

We selected the most frequent tools used for measuring psychological and psychiatric symptoms in PD, PSP, and MSA, including depression, anxiety, sleep disorder, and fatigue. The most frequent tools for depression assessment found were the Beck Depression Inventory (BDI; Beck et al., 1987), the Geriatric Depression Scale (GDS; Yesavage et al., 1982), the Hospital Anxiety and Depression Scale (HADS; Zigmond and Snaith, 1983), the Hamilton Depression Rating Scale (HAM-D; Hamilton, 1960), and the Montgomery and Asberg Depression Rating Scale (MADRS; Neumann and Schulte, 1989). Depressive symptoms consistently explained a large amount of the variance in HRQoL, regardless of the assessment tool used. For example, Santos-García and De La Fuente-Fernández (2013) found that the BDI was a good predictor of the PDQ-39 summary index, demonstrating a negative impact of depression on HRQoL (e.g., Kadastik-Eerme et al., 2015; Schrag et al., 2006; Winter et al., 2010a,b).

Regarding anxiety, several studies reported the use of the Beck Anxiety Inventory (BAI; Beck et al., 1993), HADS, and the Hamilton Anxiety Rating Scale (HAM-A; Hamilton, 1959). Similar to depression, anxiety was found to negatively impact HRQoL in PD and MSA populations (Fan et al., 2016; Meng et al., 2022; Du et al., 2018; Kovács et al., 2016). However, no information was found on anxiety's impact on HRQoL in the PSP population.

For sleep disorder evaluation, the most commonly used tools were the Epworth Sleepiness Scale (ESS; Johns, 1991) and the Parkinson's Disease Sleep Scale (PDSS; Chaudhuri et al., 2002). Although sleep disorders are addressed in NMS assessments, these two tests are specifically designed to assess sleep disorders. In the PD population, PDSS outcomes showed a negative relationship with HRQoL (Kovács et al., 2022; Liguori et al., 2021; Kovács et al., 2016; Kwon et al., 2016).

Moreover, high scores on the ESS have been shown to predict worse HRQoL outcomes in both PSP and PD populations (Shafazand et al., 2017; Kwon et al., 2013; Duncan et al., 2014; Li et al., 2023). However, no studies were found that examine the predictive value of sleep symptoms on HRQoL in MSA populations. Finally, fatigue is frequently assessed using the Fatigue Severity Scale (FSS; Krupp et al., 1989). The literature shows that higher FSS scores are associated with poorer HRQoL scores (Qin et al., 2009; Gallagher et al., 2010; Dogan et al., 2015; Sanchez-Luengos et al., 2022).

4 Discussion

The HRQoL is a multidimensional construct that reflects how individuals perceive their overall quality of life in relation to their health status. Measuring HRQoL is crucial for evaluating disease progression, assessing the effects of treatments, and understanding how specific symptoms impact an individual's quality of life. Therefore, it is essential that HRQoL measures are both valid and reliable. These measures should demonstrate clear psychometric properties, including a replicable latent structure across different groups (e.g., different countries, genders, etc.) and strong reliability indices (e.g., Cronbach's alpha).

In our study, we identified the most frequent tools used for assessing HRQoL across three different populations (PD, PSP, and MSA). We analyzed the psychometric properties of each tool. Furthermore, given the extensive number of studies that examined potential predictors of HRQoL, we described the most commonly used measures of these key predictors to highlight and address a potential issue of redundancy in this section.

4.1 Disease-specific measures

The most frequently used specific tools for assessing HRQoL in PD, PSP, and MSA are the PDQ-39, PDQ-8, PDQL, and SCOPA-PS.

4.1.1 Reliability

Several records have reported generally good reliability for these instruments, but some criticisms have been raised about the procedures used to assess them. We found extensive use of Cronbach's alpha, a useful index for evaluating internal consistency, as evidence of reliability. However, the application of Cronbach's alpha requires the measurement model to be tau-equivalent and unidimensional (Flake et al., 2017). Nonetheless, no study has ever checked for tau-equivalence in any of these specific instruments. As a result, the interpretation of Cronbach's alpha may be biased due to the lack of tau-equivalence. Moreover, we found several instances where Cronbach's alpha was applied to the summary index of the PDQ-39, despite its measurement model being multi-dimensional. In conclusion, the reliability of these specific instruments remains uncertain, and the likely violation of the assumptions required for Cronbach's alpha means we cannot draw firm conclusions regarding the reliability of the PDQ-39, PDQ-8, PDQL, and SCOPA-PS.

4.1.2 Construct validity 1: latent structure

We found different methods that scholars performed to determine the latent structure of these specific instruments, with principal component analysis (PCA), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) being the most frequently used. However, the literature reviewed presents inconsistent results, suggesting that the composition of subscales and the summary index of each instrument may be inappropriate. Variations in individuals' sociodemographic and cultural characteristics might explain these discrepancies. However, it is difficult to pinpoint which specific features are involved, as we did not find any records where ME/I was evaluated for these instruments across different groups.

4.1.3 Construct validity 2: divergent and convergent validity

In the reviewed literature, scholars have assessed divergent and convergent validity by analyzing the correlation between tests that are theoretically designed to measure the same construct (convergent validity) or distinct constructs (divergent validity). The correlation matrices generally showed good convergent validity, with several studies reporting high correlations between specific instruments and generic ones, as well as among the specific instruments themselves. However, the PDQL social subscale showed low convergent validity, and determining the cause is difficult due to the lack of information on the latent structure and the reliability of the PDQL (de Boer et al., 1996).

However, the divergent validity of these specific instruments appears problematic, as they show high correlations with tests that assess depression, anxiety, motor symptoms (MS), and non-motor symptoms (NMS). The lack of divergent validity among these specific instruments and tests used as predictors of HRQoL is concerning, as it could lead to biased interpretations in predictive studies (Serrano-Dueñas et al., 2004; Gallagher et al., 2010).

4.1.4 Recommendations

The specific instruments used to assess HRQoL in PD, PSP, and MSA could be useful tools for evaluating treatment efficacy, disease progression, and understanding how individual factors influence QoL. However, due to the limited information available on the validity and reliability of these tools, their outcomes should be interpreted with caution.

4.2 Generic instruments

The most frequently used generic tools for HRQoL assessment are the SF-36, SF-12, EQ-5D, and the NHP. These generic instruments are widely used across different populations. This discussion focuses on the psychometric characteristics of these instruments concerning their validity and reliability in PD, PSP, and MSA populations, aligning with the aim of the review. However, the following discussion section focuses only on the PD population, as validation studies of generic instruments specifically within PSP and MSA populations were not found.

4.2.1 Reliability

The SF-36 and SF-12 generally demonstrate good reliability; however, as with the specific instruments, tau-equivalence has not been checked despite Cronbach's alpha being the principal reliability index reported.

However, the NHP displayed low Cronbach's alpha values in the social isolation, energy, and sleep subscales, indicating insufficient reliability. It is important to note that only one validation study of the NHP on a PD population was found, and the sample size was limited. Consequently, the reliability of the NHP requires further empirical evidence.

No records were found that assess the reliability of the EQ-5D in PD, MSA, and PSP populations despite its widespread use (Winter et al., 2011a,b; Schrag et al., 2006; Xiao et al., 2022). This lack of reliable data suggests that the use of EQ-5D in these populations may be biased.

4.2.2 Construct validity 1: latent structure

The generic instruments discussed are self-report tools validated primarily on healthy populations, with their latent structures well-established and widely debated in the literature. However, we examined whether these latent structures remain invariant across healthy individuals and those with PD, PSP, and MSA. The invariance of latent structure is not guaranteed, and its absence could introduce significant measurement bias. We have not delved into the consequences of a lack of measurement invariance (ME/I) in various hierarchical constrained models (for more details, see Gregorich, 2006) because there are no studies addressing this issue in relation to the generic instruments. Therefore, the application of these tools to populations with PD, MSA, and PSP may be biased.

During the adaptation of tests, it could be necessary to change items significantly due to differences between the populations being studied. In such cases, a new version of the test requires a validation process as if it were a new tool. Regarding generic instruments, the

few studies that evaluated the latent structure of these tests in PD, PSP, and MSA populations appear to have followed a new validation process. Only three studies have reported a CFA on SF-36, SF-12, and an application of Item Response Theory on SF-12 in the PD population (Banks and Martin, 2009; Hagell and Westergren, 2011; Jakobsson et al., 2012). The first study shows a partial replication of SF-36/s latent structure, the second suggests a mediocre fit for SF-12/s latent model, and the third indicates a misfit for some items in SF-12. These outcomes suggest that more research is needed to evaluate the psychometric properties of these instruments in PD, PSP, and MSA populations.

4.2.3 Construct validity 2: divergent and convergent validity

We do not present knowledge about the divergent and convergent validity of generic instruments since it is consistent with the findings of the preceding section on the convergent and divergent validity of specific instruments.

4.2.4 Recommendations

General tools for HRQoL evaluation are frequently utilized across various research paradigms, including follow-up and prediction studies based on comorbidities and symptom scales as predictors. However, we observed a major issue concerning the lack of information, particularly concerning ME/I. As a result, the findings from studies that utilize generic instruments to measure HRQoL scores should be interpreted with extreme caution.

4.3 Prediction models

While summarizing and analyzing predictors of HRQoL is valuable, this section focuses on whether the tools used to assess predictors are sufficiently distinct from those used to evaluate HRQoL. Through a qualitative analysis of the scales' content, we identified a possible redundancy issue that could affect some prediction models.

4.3.1 MDS-UPDRS

In several studies, prediction models (e.g., through linear regression) have used the total score or specific subscales of the MDS-UPDRS as predictors of HRQoL. The self-report items of the MDS-UPDRS closely resemble those of the PDQ-39 and PDQL, raising concerns about redundancy when the total UPDRS score is used in prediction models. This overlap could explain the significance of the predictor. However, results from models that use only the UPDRS clinical interview sub-score suggest that these scales are significant predictors. It is likely that the objective symptoms evaluated through the UPDRS clinical interviews are good predictors of HRQoL, but the impact of the composite score should be interpreted with caution.

4.3.2 NMS-Quest and NMSS

The Non-Motor Symptoms Scale (NMSS) is frequently identified as a significant predictor of HRQoL in various studies. However, we suspect a redundancy issue here as well, since the NMSS includes items that are very similar to those in specific and generic HRQoL rating tests. For example, Rosqvist et al. (2021) found that all NMSS subscales, including mood and attention, which contain items similar to those in the PDQ-8, were significant predictors of PDQ-8 outcomes. Similar results are reported in studies by Bugalho et al. (2021), Gan et al. (2014), and Li et al., 2010.

4.3.3 H&Y

This scale is a significant predictor of HRQoL, reinforcing the idea that MSs are strong predictors of HRQoL in patients. Furthermore, the correlation between HRQoL scores and the severity of MS as rated by the H&Y scale supports the validity of HRQoL assessments, as the findings consistently show that HRQoL declines with increasing disease severity.

4.3.4 CISI-PD

The subscales of the Clinical Impression of Severity Index for Parkinson's Disease (CISI-PD) are often reported as significant predictors of HRQoL. While the subscales for motor signs, cognitive status, and motor complications are conceptually distinct from HRQoL test content, the disability subscale is quite similar to specific instruments used for HRQoL assessment. The disability subscale measures impairment in activities of daily living, as rated by a clinician. Therefore, when interpreting prediction models that include the disability subscale, caution is advised, as there may be overlap with the constructs measured by HRQoL instruments.

4.4 Comorbidities

4.4.1 Depression

Depressive symptoms are consistently associated with lower HRQoL scores, a finding that has been replicated across several studies using different tools and populations. However, these results should be interpreted with caution, as some studies do not support the divergent validity between HRQoL instruments and depression assessment tools.

Indeed, many specific and generic HRQoL tools include subscales related to "mental" wellbeing, where the items often closely resemble those found in depression assessments. This overlap raises concerns about the distinctiveness of the constructs being measured. In addition, some studies suggest that when controlling for mental HRQoL subscales, the statistical significance of depression as a predictor changes. For example, Sanchez-Luengos et al. (2022) conducted a multiple regression analysis predicting PDQ-39 outcomes using numerous variables, including the HADS. After removing the emotional wellbeing subscale, depression was no longer a significant predictor, suggesting an overlap between these features.

4.4.2 Anxiety

Anxiety is negatively correlated with HRQoL, but the lack of divergent validity between anxiety tests and HRQoL instruments suggests that these results should be interpreted carefully. Furthermore, the statistical significance of anxiety as a predictor of HRQoL appears to be dependent on the assessment tools used (Kovács et al., 2016).

4.4.3 Sleep disorder

Daily sleepiness and other sleep-related symptoms (e.g., insomnia) are associated with lower HRQoL scores, regardless of whether generic or specific instruments are administered. However, some studies have found no significant effect of sleep-related symptoms on HRQoL (Gallagher et al., 2010; Dogan et al., 2015). The lack of replication of sleep disorder effects may be related to issues with validity, reliability, and the lack of standardized cross-cultural adaptation processes for HRQoL tests, as discussed in the previous section.

4.4.4 Fatigue

Despite some items in specific instruments, such as the mobility subscale in PDQ-39, being similar to those in the Fatigue Severity Scale (FSS), FSS consistently predicts HRQoL across different assessment tools. This suggests that increased fatigue significantly worsens HRQoL. However, this result is not supported by Tu et al. (2017a,b), which may be due to the same controversies in measurement and psychometric validation discussed in the previous sections.

4.5 Implication of PD interventions

Many non-pharmacological interventions for individuals with PD include psychotherapy and physical activity (Zarotti et al., 2021; Lorenzo-García et al., 2023). In the context of psychotherapy, regardless of the core theoretical approach (e.g., cognitive, behavioral, etc.), the assessment phase is crucial for two reasons: it guides the selection of the most appropriate protocol and evaluates the efficacy of the intervention after interventions.

Cognitive Behavioral Therapy (CBT) has been shown to have a positive impact on general HRQoL in PD individuals (Berardelli et al., 2015), although there are conflicting results regarding the generalizability of its effectiveness (Zarotti et al., 2021). Similarly, the literature on mindfulness interventions presents mixed findings—some studies support their effectiveness (Vandenberg et al., 2019), while others report no significant improvements in HRQoL for PD patients (Zarotti et al., 2021).

Studies on acceptance and commitment therapy (ACT) show a similar pattern, with some improvements in emotional wellbeing but no significant changes in other HRQoL subdomains (Ghielen et al., 2017; Zarotti et al., 2021). Finally, physical activities such as dancing, Tai-chi, yoga, and Qi-Gong appear to enhance general HRQoL (Lorenzo-García et al., 2023).

Our findings suggest that these discrepancies in reported effectiveness may be partly due to potential biases in the measurement tools used to evaluate non-pharmacological therapies. The principal measures used in these evaluations (e.g., PDQ-39, ESS, FSS; Zarotti et al., 2021) may be biased due to a lack of robust psychometric information or questionable reliability assessment practices. Therefore, caution is advised when interpreting the efficacy of these interventions.

5 Limitations

Despite the lack of psychometric information being an objective problem, the issue of redundancy emerged from a qualitative analysis of the content of HRQoL instruments and predictors' tests. While some studies support the existence of redundancy, a meta-analytic approach is required to clarify the specific effects of comorbidities and symptoms on HRQoL.

6 Conclusion

HRQoL is an important construct that provides valuable insights, including the fluctuations in wellbeing experienced by patients across different contexts and stages of disease progression. Consequently, the development of accurate and reliable methods for assessing HRQoL is essential for understanding how to enhance it. However, the literature analyzed in this study reveals significant gaps in the cross-cultural validation of HRQoL evaluation techniques in PD, PSP, and MSA, as well as a general lack of in-depth investigation into their psychometric properties.

One of the critical issues identified is the absence of studies exploring measurement invariance, which is essential for ensuring that HRQoL measurements are applicable across different cultural contexts and population groups. This gap raises concerns about the generalizability of research findings and the effectiveness of cultural adaptation processes for these measurement tools. The study also highlights that many HRQoL tests validated for PD populations are being applied to PSP and MSA populations, despite the lack of literature addressing ME/I across these groups. This limitation further challenges the generalizability of the findings. Another concern is the potential redundancy in the link between HRQoL measurements and the clinical aspects of PD, PSP, and MSA. Divergent validity paradigms show that the constructs designated as HRQoL subdomains are not always clearly distinguished from other psychological features, leading to poor convergence validity. This redundancy negatively impacts the criteria validity of HRQoL measures, suggesting that the variance explained by clinical characteristics might, in fact, be due to these measures evaluating overlapping constructs.

In conclusion, while HRQoL measurements are valuable tools for assessing daily living impairments and the potential influence of therapies in the clinical populations studied, there is a pressing need for further research to address the measurement components. Strengthening the psychometric properties, especially in terms of cross-cultural validity and ME/I, is essential for ensuring that these tools provide accurate and generalizable insights into the quality of life for patients with PD, PSP, and MSA.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MM: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. RG: Data curation, Methodology, Writing – original draft, Writing – review & editing, Investigation, Formal analysis. MV: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Investigation, Project administration, Supervision, Validation, Visualization. AF: Conceptualization, Supervision, Writing – original draft. AnQ: Conceptualization, Visualization, Writing – original draft. AlQ: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the Next Generation EU—Italian NRRP, Mission 4, Component 2, Investment 1.5, under the call for the creation and strengthening of "Innovation Ecosystems," building "Territorial R&D Leaders" (Directorial Decree No. 2021/3277)—project Tech4You—Technologies for climate change adaptation and quality of life improvement, No. ECS0000009.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

These views and opinions expressed in this study are solely those of the authors, and neither the Ministry for University and Research nor the European Commission can be held responsible for them.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1438830/full#supplementary-material

References

Ader, D. N. (2007). Developing the patient-reported outcomes measurement information system (PROMIS). *Med. Care* 45, S1–S2. doi: 10.1097/01.mlr.0000260537.45076.74

Akbar, U., He, Y., Dai, Y., Hack, N., Malaty, I., McFarland, N. R., et al. (2015). Weight loss and impact on quality of life in Parkinson's disease. *PLoS ONE* 10:e0124541. doi: 10.1371/journal.pone.0124541

Alvarado-Bolaños, A., Cervantes-Arriaga, A., Rodríguez-Violante, M., Llorens-Arenas, R., Calderón-Fajardo, H., Millán-Cepeda, R., et al. (2015). Convergent validation of EQ-5D-5L in patients with Parkinson's disease. *J. Neurol. Sci.* 358, 53–57. doi: 10.1016/j.jns.2015.08.010

Antonovsky, A. (1972). Breakdown: a needed fourth step in the conceptual armamentarium of modern medicine. *Soc. Sci. Med.* 6, 537–544. doi: 10.1016/0037-7856(72)90070-4

Arboleda-Montealegre, G. Y., Cano-de-la-Cuerda, R., Fernández-de-las-Peñas, C., Sanchez-Camarero, C., and Ortega-Santiago, R. (2021). Drooling, swallowing difficulties and health related quality of life in parkinson's disease patients. *Int. J. Environ. Res. Public Health* 18:8138. doi: 10.3390/ijerph18158138

Banks, P., and Martin, C. R. (2009). The factor structure of the SF-36 in Parkinson's disease. *J. Eval. Clin. Pract.* 15, 460–463. doi: 10.1111/j.1365-2753.2008.01036.x

Beck, A. T., Epstein, N., Brown, G., and Steer, R. (1993). "*Beck anxiety inventory," in Journal of consulting and clinical psychology.*

Beck, A. T., Steer, R. A., and Brown, G. K. (1987). *Beck Depression Inventory*. New York: Harcourt Brace Jovanovich.

Benjamin, P., and Looby, J. (1998). Defining the nature of spirituality in the context of Maslow's and Rogers's theories. *Couns. Values* 42, 92–100. doi: 10.1002/j.2161-007X.1998.tb00414.x

Berardelli, I., Pasquini, M., Bloise, M., Tarsitani, L., Biondi, M., Berardelli, A., et al. (2015). CBT group intervention for depression, anxiety, and motor symptoms in Parkinson's disease: preliminary findings. *Int. J. Cogn. Ther.* 8, 11–20. doi: 10.1521/ijct.2015.8.1.11

Berg, K., Wood-Dauphinee, S., and Williams, J. I. (1995). The balance scale: reliability assessment with elderly residents and patients with an acute stroke. *Scand. J. Rehabil. Med.* 27, 27–36.

Bergner, M., Bobbitt, R. A., Pollard, W. E., Martin, D. P., and Gilson, B. S. (1976). The sickness impact profile: validation of a health status measure. *Med. Care* 14, 57–67. doi: 10.1097/00005650-197601000-00006

Biggs, J. T., Wylie, L. T., and Ziegler, V. E. (1978). Validity of the Zung self-rating depression scale. *Br. J. Psychiatry* 132, 381–385 doi: 10.1192/bjp.132.4.381

Boxer, A. L., Yu, J.-T., Golbe, L. I., Litvan, I., Lang, A. E., and Höglinger, G. U. (2017). Advances in progressive supranuclear palsy: new diagnostic criteria, biomarkers, and therapeutic approaches. *Lancet Neurol.* 16, 552–563. doi: 10.1016/s1474-4422(17)30157-6

Brazier, J., Roberts, J., and Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *J. Health Econ.* 21, 271–292. doi: 10.1016/S0167-6296(01)00130-8

Brown, C. A., Cheng, E. M., Hays, R. D., Vassar, S. D., and Vickrey, B. G. (2009). SF-36 includes less Parkinson Disease (PD)-targeted content but is more responsive to

change than two PD-targeted health-related quality of life measures. *Qual. Life Res.* 18, 1219–1237. *doi: 10.1007/s11136-009-9530-y*

Brown, K. W., and Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological wellbeing. *J. Pers. Soc. Psychol.* 84:822. doi: 10.1037/0022-3514.84.4.822

Brown, R. G., Dittner, A., Findley, L., and Wessely, S. C. (2005). The Parkinson fatigue scale. *Parkins. Related Disor.* 11, 49–55. doi: 10.1016/j.parkreldis.2004.07.007

Bugalho, P., Ladeira, F., Barbosa, R., Marto, J. P., Borbinha, C., da Conceição, L., et al. (2021). Progression in Parkinson's disease: variation in motor and non-motor symptoms severity and predictors of decline in cognition, motor function, disability, and health-related quality of life as assessed by two different methods. *Move. Disor. Clin. Pract.* 8, 885–895. doi: 10.1002/mdc3.13262

Buysse, D. J., Reynolds, 3rd, C. F., Monk, T. H., Berman, S. R., and Kupfer, D.J. (1989). The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatr. Res.* 28, 193–213. doi: 10.1016/0165-1781(89)90047-4

Cardol, M., de Haan, R. J., de Jong, B. A., Van den Bos, G. A., and de Groot, I. J. (2001). Psychometric properties of the impact on participation and autonomy questionnaire. *Arch. Phys. Med. Rehabil.* 82, 210–216. doi: 10.1053/apmr.2001.18218

Chan, C. H., Chan, T. H., Leung, P. P., Brenner, M. J., Wong, V. P., Leung, E. K., et al. (2014). Rethinking wellbeing in terms of affliction and equanimity: development of a holistic wellbeing scale. *J. Ethnic Cult. Diver. Soc. Work* 23, 289–308. doi: 10.1080/15313204.2014.932550

Chaudhuri, K. R., Martinez-Martin, P., Brown, R. G., Sethi, K., Stocchi, F., Odin, P., et al. (2007). The metric properties of a novel non-motor symptoms scale for Parkinson's disease: results from an international pilot study. *Mov. Disord.* 22, 1901–1911. doi: 10.1002/mds.21596

Chaudhuri, K. R., Pal, S., DiMarco, A., Whately-Smith, C., Bridgman, K., Mathew, R., et al. (2002). The Parkinson's disease sleep scale: a new instrument for assessing sleep and nocturnal disability in Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 73, 629–635. doi: 10.1136/jnnp.73.6.629

Chaudhuri, K. R., Rizos, A., Trenkwalder, C., Rascol, O., Pal, S., Martino, D., et al. (2015). King's Parkinson's disease pain scale, the first scale for pain in PD: an international validation. *Movem. Disor.* 30, 1623–1631. doi: 10.1002/mds.26270

Chaudhuri, K. R., Rojo, J. M., Schapira, A. H., Brooks, D. J., and Stocchi, F., Odin, et al. (2013). A proposal for a comprehensive grading of Parkinson's disease severity combining motor and non-motor assessments: meeting an unmet need. *PLoS ONE* 8:e57221. doi: 10.1371/journal.pone.0057221

Chen, K., Yang, Y. J., Liu, F. T., Li, D. K., Bu, L. L., Yang, K., et al. (2017). Evaluation of PDQ-8 and its relationship with PDQ-39 in China: A three-year longitudinal study. *Health Qual. Life Outc.* 15. doi: 10.1186/s12955-017-0742-5

Cherbuin, N., and Francis Jorm, A. (2010). "The informant Questionnaire on cognitive decline in the elderly (IQCODE)," in *Principles and Practice of Geriatric Psychiatry*, 147–151. doi: 10.1002/9780470669600.ch28

Chogahara, M. (1999). A multidimensional scale for assessing positive and negative social influences on physical activity in older adults. *J. Gerontol. Series B* 54, S356–S367. doi: 10.1093/geronb/54B.6.S356

Chuquilín-Arista, F., Álvarez-Avellón, T., and Menéndez-González, M. (2021). Impact of depression and anxiety on dimensions of health-related quality of life in subjects with parkinson's disease enrolled in an association of patients. *Brain Sci.* 11:771. doi: 10.3390/brainsci11060771

de Boer, A. G., Wijker, W., Speelman, J. D., and de Haes, J. C. (1996). Quality of life in patients with Parkinson's disease: development of a questionnaire. *J. Neurol. Neurosurg. Psychiatr.* 61, 70–74. doi: 10.1136/jnnp.61.1.70

DeYoung, C. G., and Lena, C. Q., and Jordan, B. P. (2007). Between facets and domains: 10 aspects of the Big Five. *J. Pers. Soc. Psychol.* 93:880. doi: 10.1037/0022-3514.93.5.880

D'Iorio, A., Vitale, C., Piscopo, F., Baiano, C., Falanga, A. P., Longo, K., et al. (2017). Impact of anxiety, apathy and reduced functional autonomy on perceived quality of life in Parkinson's disease. *Parkinson. Relat. Disor.* 43, 114–117. doi: 10.1016/j.parkreldis.2017.08.003

Dodds, A. G., Bailey, P., Pearson, A., and Yates, L. (1991). Psychological factors in acquired visual impairment: the development of a scale of adjustment. *J. Visual Impair. Blind.* 85, 306–310. doi: 10.1177/0145482X9108500711

Dogan, V. B., Koksal, A., Dirican, A., Baybas, S., Dirican, A., and Dogan, G. B. (2015). Independent effect of fatigue on health-related quality of life in patients with idiopathic Parkinson's disease. *Neurol. Sci.* 36, 2221–2226. doi: 10.1007/s10072-015-2340-9

Du, J. J., Wang, T., Huang, P., Cui, S., Gao, C., Lin, Y., et al. (2018). Clinical characteristics and quality of life in Chinese patients with multiple system atrophy. *Brain Behav.* 8:e01135. doi: 10.1002/brb3.1135

Dubois, B., Burn, D., Goetz, C., Aarsland, D., Brown, R. G., Broe, G. A., et al. (2007). Diagnostic procedures for Parkinson's disease dementia: recommendations from the movement disorder society task force. *Move. Disor.* 22, 2314–2324. doi: 10.1002/mds.21844

Duncan, G. W., Khoo, T. K., Yarnall, A. J., O'Brien, J. T., Coleman, S. Y., Brooks, D. J., et al. (2014). Health-related quality of life in early Parkinson's disease: The impact of nonmotor symptoms. *Move. Disor.* 29, 195–202. doi: 10.1002/mds.25664

Endler, N., and Parker, J. D. (1999). *Coping inventory for stressful situations*. doi: 10.1037/t67919-000

Euroqol Group (1990). EuroQol-a new facility for the measurement of health-related quality of life. *Health Policy* 16, 199–208. doi: 10.1016/0168-8510(90)90421-9

Fahn, S. R. L. E. (1987). "Unified Parkinson's disease rating scale," in *Recent Developments in Parkinson's Disease* 153–163.

Fan, J. Y., Chang, B. L., and Wu, Y. R. (2016). Relationships among depression, anxiety, sleep, and quality of life in patients with Parkinson's disease in Taiwan. *Parkinson's Dis.* 2016:4040185. doi: 10.1155/2016/4040185

Fereshtehnejad, S. M., Farhadi, F., Hadizadeh, H., Shahidi, G. A., Delbari, A., and Lökk, J. (2014a). Cross-cultural validity, reliability, and psychometric properties of the persian version of the scales for outcomes in parkinson's disease-psychosocial questionnaire. *Neurol. Res. Int.* 2014:260684. doi: 10.1155/2014/260684

Fereshtehnejad, S. M., Naderi, N., Rahmani, A., Shahidi, G. A., Delbari, A., and Lökk, J. (2014b). Psychometric study of the Persian short-form eight-item Parkinson's disease questionnaire (PDQ-8) to evaluate health related quality of life (HRQoL). *Health Qual. Life Outc.* 12, 1–9. doi: 10.1186/1477-7525-12-78

Ferguson, L., and Scheman, J. (2009). Patient global impression of change scores within the context of a chronic pain rehabilitation program. *J. Pain* 10:S73. doi: 10.1016/j.jpain.2009.01.258

Fillenbaum, G. G., and Smyer, M. A. (1981). The development, validity, and reliability of the OARS multidimensional functional assessment questionnaire. *J. Gerontol.* 36, 428–434. doi: 10.1093/geronj/36.4.428

Flake, J. K., Pek, J., and Hehman, E. (2017). Construct validation in social and personality research: current practice and recommendations. *Soc. Psychol. Personal. Sci.* 8, 370–378. doi: 10.1177/1948550617693063

Folkman, S., and Lazarus, R. S. (1985). If it changes it must be a process: study of emotion and coping during three stages of a college examination. *J. Pers. Soc. Psychol.* 48:150. doi: 10.1037//0022-3514.48.1.150

Franchignoni, F., Giordano, A., and Ferriero, G. (2008). Rasch analysis of the short form 8-item Parkinson's disease questionnaire (PDQ-8). *Qual. Life Res.* 17, 541–548. doi: 10.1007/s11136-008-9341-6

Franke, G. H., Mähner, N., Reimer, J., Spangemacher, B., and Esser, J. (2000). *ErsteÜberprüfung des Essener Fragebogenszur Krankheitsverarbeitung (EFK) ansehbeeinträchtigten Patienten*. London: Kudos Innovations Ltd. doi: 10.1024//0170-1789.21.2.166

Franke, G. H.;, Nentzl, J., and Jagla-Franke, M. (2020). *SAMS. Stendal Adherence to Medication Score*. Available online: https://www.psychometrikon.de/inhalt/suchen/test.php?id=ff32ee9ea015021c3fb047e505e2bc45 (accessed February 22, 2021).

Furlong, W., Feeny, D., Torrance, G., Goldsmith, C., DePauw, S., and Zhu, Z., et al. (1998). *Multiplicative multi-attribute utility function for the Health Utilities Index Mark 3 (HUI3) system: a technical report* (No. 1998-11). Centre for Health Economics and Policy Analysis (CHEPA), McMaster University, Hamilton, Canada.

Galeoto, G., Colalelli, F., Massai, P., Berardi, A., Tofani, M., Pierantozzi, M., et al. (2018). Quality of life in Parkinson's disease: Italian validation of the Parkinson's Disease Questionnaire (PDQ-39-IT). *Neurol. Sci.* 39, 1903–1909. doi: 10.1007/s10072-018-3524-x

Gallagher, D. A., Lees, A. J., and Schrag, A. (2010). What are the most important nonmotor symptoms in patients with Parkinson's disease and are we missing them? *Move. Disor.* 25, 2493–2500. doi: 10.1002/mds.23394

Gan, J., Zhou, M., Chen, W., and Liu, Z. (2014). Non-motor symptoms in Chinese Parkinson's disease patients. *J. Clin. Neurosci.* 21, 751–754. doi: 10.1016/j.jocn.2013.07.015

Gershon, R., Lai, J., Bode, R., Choi, S., Moy, C., Bleck, T., et al. (2012). Neuro-QOL: quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing. *Q. Life Res.* 21, 475–486. doi: 10.1007/s11136-011-9958-8

Ghielen, I., van Wegen, E. E., Rutten, S., de Goede, C. J., Houniet-de Gier, M., Collette, E. H., et al. (2017). Body awareness training in the treatment of wearing-off related anxiety in patients with Parkinson's disease: Results from a pilot randomized controlled trial. *J. Psychosom. Res.* 103, 1–8. doi: 10.1016/j.jpsychores.2017.09.008

Giladi, N., Shabtai, H., Simon, E. S., Biran, S., Tal, J., and Korczyn, A. D. (2000). Construction of freezing of gait questionnaire for patients with Parkinsonism. *Parkins. Relat. Disor.* 6, 165–170. doi: 10.1016/S1353-8020(99)00062-0

Global Parkinson's Disease Survey Steering Committee (2002). Factors impacting on quality of life in Parkinson's disease: results from an international survey. *Mov Disord.* 17, 60–67. doi: 10.1002/mds.10010

Goetz, C. G., Stebbins, G. T., Shale, H. M., Lang, A. E., Chernik, D. A., Chmura, T. A., et al. (1994). Utility of an objective dyskinesia rating scale for Parkinson's disease: inter- and intrarater reliability assessment. *Move. Disor.* 9, 390–394. doi: 10.1002/mds.870090403

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., et al. (2008). Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Move. Disor.* 23, 2129–2170. doi: 10.1002/mds.22340

Golbe, L. I., and Ohman-Strickland, P. A. (2007). A clinical rating scale for progressive supranuclear palsy. *Brain J. Neurol.* 130, 1552–1565. doi: 10.1093/brain/awm032

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? *Testing measurement invariance using the confirmatory factor analysis framework. Med. Care* 44, S78–S94. doi: 10.1097/01.mlr.0000245454.12228.8f

Grimbergen, Y. A., Schrag, A., Mazibrada, G., Borm, G. F., and Bloem, B. R. (2013). Impact of falls and fear of falling on health-related quality of life in patients with Parkinson's disease. *J. Parkinson's Dis.* 3, 409–413. doi: 10.3233/JPD-120113

Guo, X., Song, W., Chen, K., Chen, X., Zheng, Z., Cao, B., et al. (2015). Impact of Frontal Lobe Function and Behavioral Changes on Health-Related Quality of Life in Patients with Parkinson's Disease: A Cross-Sectional Study from Southwest China. *EuropeanNeurology* 74, 147–153. doi: 10.1159/000439084

Hagell, P., and Westergren, A. (2011). Measurement properties of the SF-12 health survey in Parkinson's disease. *J. Parkinson's Dis.* 1, 185–196. doi: 10.3233/JPD-2011-11026

Hagell, P., Whalley, D., McKenna, S. P., and Lindvall, O. (2003). Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham Health Profile. *Mov. Disord.* 18, 773–778. doi: 10.1002/mds.10438

Hamilton, M. (1959). The assessment of anxiety states by rating. *Br. J. Med. Psychol.* 32, 50–55. doi: 10.1111/j.2044-8341.1959.tb00467.x

Hamilton, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatr.* 23:56. doi: 10.1136/jnnp.23.1.56

Hechtner, M. C., Vogt, T., Zöllner, Y., Schröder, S., Sauer, J. B., Binder, H., et al. (2014). Quality of life in Parkinson's disease patients with motor fluctuations and dyskinesias in five European countries. *Parkins. Relat. Disor.* 20, 969–974. doi: 10.1016/j.parkreldis.2014.06.001

Higgins, J. P., and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*. 4th ed. Chichester: John Wiley and Sons.

Hoehn, M. M., and Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology* 17, 427–442. doi: 10.1212/WNL.17.5.427

Hoehn, M. M., and Yahr, M. D. (1998). Parkinsonism: onset, progression, and mortality. *Neurology* 50:318. doi: 10.1212/WNL.50.2.318

Horak, F. B., Wrisley, D. M., and Frank, J. (2009). The balance evaluation systems test (BESTest) to differentiate balance deficits. *Phys. Ther.* 89, 484–498. doi: 10.2522/ptj.20080071

Hunt, S. M., McEwen, J., and McKenna, S. P. (1985). Measuring health status: a new tool for clinicians and epidemiologists. *J. R. Coll. Gen. Pract.* 35, 185–188.

Hurst, H., and Bolton, J. (2004). Assessing the clinical significance of change scores recorded on subjective outcome measures. *J. Manipulative Physiol. Ther.* 27, 26–35. doi: 10.1016/j.jmpt.2003.11.003

Jahanshahi, M., and Marsden, C. D. (1988). Personality in torticollis: a controlled study. *Psychol. Med.* 18:375–387. doi: 10.1017/S0033291700007923

Jakobsson, U., Westergren, A., Lindskov, S., and Hagell, P. (2012). Construct validity of the SF-12 in three different samples. *J. Eval. Clin. Pract.* 18, 560–566. doi: 10.1111/j.1365-2753.2010.01623.x

Jecmenica-Lukic, M. V., Pekmezovic, T. D., Petrovic, I. N., Dragasevic, N. T., and Kosti,ć, V. S. (2018). Factors associated with deterioration of health-related quality of life in multiple system atrophy: 1-year follow-up study. *Acta NeurologicaBelgica* 118, 589–595. doi: 10.1007/s13760-018-0962-4

Jenkins, C. D., Jono, R. T., Stanton, B. A., and Stroup-Benham, C. A. (1990). The measurement of health-related quality of life: major dimensions identified by factor analysis. *Soc. Sci. Med.* 31, 925–931. doi: 10.1016/0277-9536(90)90032-N

Jenkinson, C., Fitzpatrick, R., Peto, V., Greenhall, R., and Hyman, N. (1997). The PDQ-8: development and validation of a short-form Parkinson's disease questionnaire. *Psychol. Health*, 12, 805–814. doi: 10.1080/08870449708406741

Jette, A. M., Davies, A. R., Cleary, P. D., Calkins, D. R., Rubenstein, L. V., Fink, A., et al. (1986). The Functional StatusQuestionnaire: reliability and validity when used in primary care. *J. Gen. Intern. Med.* 1, 143–149. doi: 10.1007/BF02602324

Johns, M. W. (1991). A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 14, 540–545. doi: 10.1093/sleep/14.6.540

Josiah, A. F., Gruber-Baldini, A. L., Anderson, K. E., Fishman, P. S., Weiner, W. J., Reich, S. G., et al. (2012). The effects of gait impairment with and without freezing of gait in Parkinson's disease. *Parkinsonism Relat. Disor.* 18, 239–242. doi: 10.1016/j.parkreldis.2011.10.008

Kadastik-Eerme, L., Rosenthal, M., Paju, T., Muldmaa, M., and Taba, P. (2015). Health-related quality of life in Parkinson's disease: A cross-sectional study focusing on non-motor symptoms. *Health Qual. Life Outcomes* 13, 1–8. doi: 10.1186/s12955-015-0281-x

Karlsen, K. H., Tandberg, E., Årsland, D., and Larsen, J. P. (2000). Health related quality of life in Parkinson's disease: a prospective longitudinal study. *J. Neurol. Neurosurg. Psychiat.* 69, 584–589. doi: 10.1136/jnnp.69.5.584

Kertesz, A., Davidson, W., and Fox, H. (1997). Frontal behavioral inventory: diagnostic criteria for frontal lobe dementia. *Canadian J. Neurol. Sci.* 24, 29–36. doi: 10.1017/S0317167100021053

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., and Wittchen, H. U. (1998). The World Health Organization composite international diagnostic interview short-form (CIDI-SF). *Int. J. Methods Psychiatr. Res.* 7, 171–185. doi: 10.1002/mpr.47

Kim, Y. E., Kim, H. J., Yun, J. Y., Lee, W. W., Yang, H. J., Kim, J. M., et al. (2018). *Musculoskeletal problems affect the quality of life of patients with Parkinson's* disease. *J. Move. Disor.* 11:133. doi: 10.14802/jmd.18022

Kovács, M., Makkos, A., Aschermann, Z., Janszky, J., Komoly, S., Weintraut, R., et al. (2016). Impact of sex on the nonmotor symptoms and the health-related quality of life in Parkinson's disease. *Parkinsons Dis.* 2016:7951840. doi: 10.1155/2016/7951840

Kovács, N., Bergmann, L., Anca-Herschkovitsch, M., Cubo, E., Davis, T. L., Iansek, R., et al. (2022). Outcomes impacting quality of life in advanced Parkinson's disease patients treated with levodopa-carbidopa intestinal gel. *J. Parkinsons Dis.* 12, 917–926. doi: 10.3233/jpd-212979

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Kroll, T., Kehn, M., Ho, P. S., and Groah, S. (2007). The SCI exercise self-efficacy scale (ESES): development and psychometric properties. *Int. J. Behav. Nutr. Phys. Act.* 4:34. doi: 10.1186/1479-5868-4-34

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., and Steinberg, A. D. (1989). The fatigue severity scale: application to patients with multiple sclerosis and systemic lupus erythematosus. *Arch. Neurol.* 46, 1121–1123. doi: 10.1001/archneur.1989.00520460115022

Kwon, D. Y., Kim, J. W., Ma, H., Il, Ahn, T. B., Cho, J., Lee, P. H., et al. (2013). Translation and validation of the Korean version of the 39-item Parkinson's disease questionnaire. *J. Clin. Neurol.* 9, 26–31. doi: 10.3988/jcn.2013.9.1.26

Kwon, D. Y., Koh, S. B., Lee, J. H., Park, H. K., Kim, H. J., and Shin, H. W. (2016). The KMDS- NATION study: Korean Movement Disorder society multicenter assessment of non-motor symptoms and quality of life in Parkinson's disease NATION study group. *J. Clin. Neurol.* 12, 393–402. doi: 10.3988/jcn.2016.12.4.393

Lau, A. L. D., Cummins, R. A., and McPherson, W. (2005). An investigation into the cross-cultural equivalence of the personal wellbeing index. *Soc. Indic. Res.* 72, 403–430. doi: 10.1007/s11205-004-0561-z

Lawton, M. P., and Brody, E. M. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontol. 9,* 179–186. doi: 10.1093/geront/9.3_Part_1.179

Lee, J. H., Choi, M. K., Jung, D., Sohn, Y. H., and Hong, J. Y. (2015). A structural model of health-related quality of life in Parkinson's disease patients. *West. J. Nurs. Res.* 37, 1062–1080. doi: 10.1177/0193945914528588

Lee, K. A., Hicks, G., and Nino-Murcia, G. (1991). Validity and reliability of a scale to assess fatigue. *Psychiat. Res.* 36, 291–298. doi: 10.1016/0165-1781(91)90027-M

Leentjens, A. F., Dujardin, K., Pontone, G. M., Starkstein, S. E., Weintraub, D., and Martinez-Martin, P. (2014). The Parkinson Anxiety Scale (PAS): development and validation of a new anxiety scale. *Move. Disor.* 29, 1035–1043. doi: 10.1002/mds.25919

Leroi, I., Ahearn, D. J., Andrews, M., Mcdonald, K. R., Byrne, E. J., and Burns, A. (2011). Behavioural disorders, disability and quality of life in Parkinson's disease. *Age Ageing* 40, 614–621. doi: 10.1093/ageing/afr078

Li, H., Zhang, M., Chen, L., Zhang, J., Pei, Z., Hu, A., et al. (2010). Nonmotor symptoms are independently associated with impaired health-related quality of life in Chinese patients with Parkinson's disease. *Move. Disor.* 25, 2740–2746. doi: 10.1002/mds.23368

Li, X. Y., Chen, M. J., Liang, X. N., Yao, R. X., Shen, B., Wu, B., et al. (2023). PDQ-8: a simplified and effective tool measuring life quality in progressive supranuclear palsy. *J. Parkinson's Dis.* 13, 83–91. doi: 10.3233/JPD-223553

Liguori, C., De Franco, V., Cerroni, R., Spanetta, M., Mercuri, N. B., Stefani, A., et al. (2021). Sleep problems affect quality of life in Parkinson's disease along disease progression. *Sleep Med.* 81, 307–311. doi: 10.1016/j.sleep.2021.02.036

Livingston, G., Blizard, B., and Mann, A. (1993). Does sleep disturbance predict depression in elderly people? *A study in inner London. Br. J. General Pract.* 43, 445–448.

Logsdon, R. G., Gibbons, L. E., McCurry, S. M., and Teri, L. (1999). Quality of life in Alzheimer's disease: patient and caregiver reports. *J. Ment. Health Aging* 5, 21–32. doi: 10.1037/t03352-000

Lorenzo-García, P., de Arenas-Arroyo, S. N., Cavero-Redondo, I., Guzmán-Pavón, M. J., Priego-Jiménez, S., and Álvarez-Bueno, C. (2023). Physical exercise interventions on quality of life in parkinson disease: a network meta-analysis. *J. Neurol. Phys. Ther.* 47, 64–74. doi: 10.1097/NPT.0000000000000414

Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with

the beck depression and anxiety inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-U

Luo, N., Tan, L. C. S., Zhao, Y., Lau, P. N., Au, W. L., and Li, S. C. (2009). Determination of the longitudinal validity and minimally important difference of the 8-item Parkinson's disease questionnaire (PDQ-8). *Move. Disor.* 24, 183–187. doi: 10.1002/mds.22240

Mahoney, F. I., and Barthel, D. W. (1965). Functional evaluation: the barthel index. *Maryland State Med. J.* 14, 61–65. doi: 10.1037/t02366-000

Main, C. J. (1983). The modified somatic perception questionnaire (MSPQ). *J. Psychosom. Res.* 27, 503–514. doi: 10.1016/0022-3999(83)90040-5

Marin, R. S., Biedrzycki, R. C., and Firinciogullari, S. (1991). Reliability and validity of the apathy evaluation scale. *Psychiatry Res.* 38, 143–162. doi: 10.1016/0165-1781(91)90040-V

Marinus, J., Visser, M., Martinez-Martin, P., van Hilten, J. J., and Stiggelbout, A. M. (2003). A short psychosocial questionnaire for patients with Parkinson's disease: the SCOPA-PS. *J. Clin. Epidemiol.* 56, 61–67. doi: 10.1016/S0895-4356(02)00569-3

Martínez-Martín, P., Benito-León, J., Alonso, F., Catalán, M. J., Pondal, M., Tobías, A., et al. (2003). Patients', doctors', and caregivers' assessment of disability using the UPDRS-ADL section: are these ratings interchangeable? *Move. Disor.* 18, 985–992. doi: 10.1002/mds.10479

Martinez-Martin, P., Forjaz, M. J., Frades, B., and De Pedro-Cuesta, J. (2005). La rapid assessment disability scale (RAD) enEnfermedad de Parkinson. *GacSanit* 19:68.

Martínez-Martín, P., Rodríguez-Blázquez, C., Forjaz, M. J., Álvarez-Sánchez, M., Arakaki, T., Bergareche-Yarza, A., et al. (2014). Relationship between the MDS-UPDRS domains and the health-related quality of life of Parkinson's disease patients. *Eur. J. Neurol.* 21, 519–524. doi: 10.1111/ene.12349

Meng, D., Jin, Z., Chen, K., Yu, X., Wang, Y., Du, W., et al. (2022). Quality of life predicts rehabilitation prognosis in Parkinson's disease patients: factors influence rehabilitation prognosis. *Brain Behav.* 12:e2579. doi: 10.1002/brb3.2579

Muthny, F. A. (1989). *Freiburger Fragebogenzur Krankheitsverarbeitung: FKV*. Weinheim: Beltz. doi: 10.1007/978-3-642-74986-5_7

Nakano, T., Kajiyama, Y., Revankar, G. S., Hashimoto, R., Watanabe, Y., Kishima, H., et al. (2021). Neural networks associated with quality of life in patients with Parkinson's disease. *Parkins. Relat. Disor.* 89, 6–12. doi: 10.1016/j.parkreldis.2021.06.007

Neumann, N. U., and Schulte, R. M. (1989). *Montgomery and Asberg Depression Rating Scale*. Deutsche Fassung. Erlangen: Perimed Fachbuch Verlagsgesellschaft.

Norlin, J. M., Kellerborg, K., Persson, U., Åström, D. O., and Hagell, P., Martinez-Martin, P., et al. (2023). Clinical impression of severity index for Parkinson's disease and its association to health-related quality of life. *Move. Disor. Clin. Practice* 10, 392–398. doi: 10.1002/mdc3.13649

Ophey, A., Eggers, C., Dano, R., Timmermann, L., and Kalbe, E. (2018). Health-related quality of life subdomains in patients with Parkinson's disease: the role of gender. *Parkinson's Disease* 2018:6532320. doi: 10.1155/2018/6532320

Ou, R., Liu, H., Hou, Y., Wei, Q., Cao, B., Zhao, B., et al. (2017). Executive dysfunction, behavioral changes and quality of life in Chinese patients with progressive supranuclear palsy. *J. Neurol. Sci.* 380, 182–186. doi: 10.1016/j.jns.2017.07.033

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021a). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* 88:105906.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021b). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372:n160. doi: 10.1136/bmj.n160

Patton, J. H., Stanford, M. S., and Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *J. Clin. Psychol.* 51, 768–774. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Peto, V., Jenkinson, C., Fitzpatrick, R., and Greenhall, R. (1995). The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease. *Qual. Life Res.* 4, 241–248. doi: 10.1007/BF02260863

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *J. Am. Geriatr. Soc.* 23, 433–441. doi: 10.1111/j.1532-5415.1975.tb00927.x

Picillo, M., Cuoco, S., Amboni, M., Bonifacio, F. P., Bruschi, F., Carotenuto, I., et al. (2019). Validation of the Italian version of the PSP Quality of Life questionnaire. *Neurol. Sci.* 40, 2587–2594. doi: 10.1007/s10072-019-04010-2

Powell, L. E., and Myers, A. M. (1995). The activities-specific balance confidence (ABC) scale. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* 50, M28–M34. doi: 10.1093/gerona/50a.1.m28

Qin, Z., Zhang, L., Sun, F., Liu, H., Fang, X., Chan, P., et al. (2009). Depressive symptoms impacting on health-related quality of life in early Parkinson's disease: results from Chinese l-dopa exposed cohort. *Clin. Neurol. Neurosurg.* 111, 733–737. doi: 10.1016/j.clineuro.2009.07.001

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. doi: 10.1177/014662167700100306

Reilly, M. C., Zbrozek, A. S., and Dukes, E. M. (1993). The validity and reproducibility of a work productivity and activity impairment instrument. *Pharmaco Econ.* 4, 353–365. doi: 10.2165/00019053-199304050-00006

Reisberg, B., Borenstein, J., Salob, S. P., Ferris, S. H., Franssen, E., and Georgotas, A. (1987). Behavioral symptoms in Alzheimer's disease: phenomenology and treatment. *J. Clin. Psychiatry* 48, 9–15. doi: 10.1037/t13385-000

Resnick, B., and Jenkins, L. S. (2000). Testing the reliability and validity of the self-efficacy for exercise scale. *Nurs. Res.* 49, 154–159. doi: 10.1097/00006199-200005000-00007

Richards, K. (1987). Techniques for measurement of sleep in critical care. *Focus Crit. Care* 14, 34–40.

Ringendahl, H., Werheid, K., Leplow, B., Ellgring, H., Annecke, R., and Emmans, D. (2000). VorschlägeFür Eine Standardisierte psychologische Diagnostik bei Parkinsonpatienten. *DerNervenarzt* 71, 946–954. doi: 10.1007/s001150050691

Roach, A. J., Frazier, L. P., and Bowden, S. R. (1981). The marital satisfaction scale: development of a measure for intervention research. *J. Marriage Fam.* 537–546. doi: 10.2307/351755

Romenets, S. R., Wolfson, C., Galatas, C., Pelletier, A., Altman, R., Wadup, L., et al. (2012). Validation of the non-motor symptoms questionnaire (NMS-Quest). *Parkins. Related Disor.* 18, 54–58. doi: 10.1016/j.parkreldis.2011.08.013

Rosenberg, M. (1965). *Rosenberg self-esteem scale. J. Relig. Health* 1:3080. doi: 10.1037/t01038-000

Rosenstiel, A. K., and Keefe, F. J. (1983). The use of coping strategies in chronic low back pain patients: relationship to patient characteristics and current adjustment. *Pain* 17, 33–44. doi: 10.1016/0304-3959(83)90125-2

Rosqvist, K., Odin, P., Lorenzl, S., Meissner, W. G., Bloem, B. R., Ferreira, J. J., et al. (2021). Factors associated with health-related quality of life in late-stage Parkinson's disease. *Move. Disor. Clin. Pract.* 8, 563–570. doi: 10.1002/mdc3.13186

Rybarczyk, B. (2011). Social and occupational functioning assessment scale (SOFAS). *Encyclopedia Clin. Neuropsychol.* 63:2313. doi: 10.1007/978-0-387-79948-3_428

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological wellbeing. *J. Pers. Soc. Psychol.* 57, 1069–1081. doi: 10.1037/0022-3514.57.6.1069

Ryff, C. D., and Keyes, C. L. M. (1995). The structure of psychological wellbeing. *J. Pers. Soc. Psychol.* 69, 719–727. doi: 10.1037/0022-3514.69.4.719

Sanchez-Luengos, I., Lucas-Jiménez, O., Ojeda, N., Peña, J., and Gómez-Esteban, J. C., Gómez-Beldarrain, et al. (2022). Predictors of health-related quality of life in Parkinson's disease: the impact of overlap between health-related quality of life and clinical measures. *Qual. Life Res.* 31, 3241–3252. doi: 10.1007/s11136-022-03187-y

Santangelo, G., Barone, P., Cuoco, S., Raimo, S., Pezzella, D., Picillo, M., et al. (2014). Apathy in untreated, de novo patients with Parkinson's disease: validation study of Apathy Evaluation Scale. *J. Neurol.* 261, 2319–2328. doi: 10.1007/s00415-014-7498-1

Santos-García, D., and De La Fuente-Fernández, R. (2013). Impact of non-motor symptoms on health-related and perceived quality of life in Parkinson's disease. *J. Neurol. Sci. 332,* 136–140. doi: 10.1016/j.jns.2013.07.005

Sarason, I. G., Levine, H. M., Basham, R. B., and Sarason, B. R. (1983). Assessing social support: the social support questionnaire. *J. Pers. Soc. Psychol.* 44, 127–139 doi: 10.1037/0022-3514.44.1.127

Savci, C., and Sendir, M. (2009). Evaluation of health related quality of life in patients with Parkinsons disease. *Neurosciences* 14, 60–66.

Scheier, M. F., and Carver, C. S. (1985). Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health Psychol.* 4:219. doi: 10.1037//0278-6133.4.3.219

Schmidt, J., Lamprecht, F., and Wittmann, W. W. (1989). Satisfaction with inpatient management. Development of a questionnaire and initial validity studies. *Psychother. Psychosom. Med. Psychol.* 39, 248–255

Schönenberg, A., Santos García, D., Mir, P., Wu, J. J., Heimrich, K. G., Mühlhammer, H. M., et al. (2023). Using network analysis to explore the validity and influential items of the Parkinson's Disease Questionnaire-39. *Sci. Rep.* 13:7221. doi: 10.1038/s41598-023-34412-4

Schrag, A., Hovris, A., Morley, D., Quinn, N., and Jahanshahi, M. (2003). Young- versus older-onset Parkinson's disease: Impact of disease and psychosocial consequences. *Move. Disor.* 18, 1250–1256. doi: 10.1002/mds.10527

Schrag, A., Selai, C., Mathias, C., Low, P., Hobart, J., Brady, N., et al. (2007). Measuring health-related quality of life in MSA: the MSA-QoL. *Movement Disor.* 22, 2332–2338. doi: 10.1002/mds.21649

Schrag, A., Selai, C., Quinn, N., Lees, A., Litvan, I., Lang, A., et al. (2006). Measuring quality of life in PSP: the PSP-QoL. *Neurology* 67, 39–44. doi: 10.1212/01.wnl.0000223826.84080.97

Schwab, R. S., and England, A. C. J. (1969). "Projection technique for evaluating surgery in Parkinson's disease," in *Third Symposium on Parkinson's Disease,* eds. F. J. Gillingham, I. M. L. Donaldson (*Edinburgh, Scotland: E and S Livingstone),* 152–157.

Schwarzer, R., and Jerusalem, M. (1999). *SkalenzurErfassung von Lehrer- und Schülermerkmalen [Scales for the Survey of Traits of Teachers and Students], Freie Universität, Berlin*. Available at: http://www.psyc.de/skalendoku.pdf (accessed July 21, 2024).

Serrano-Dueñas, M., Martínez-Martín, P., and Vaca-Baquero, V. (2004). Validation and cross-cultural adjustment of PDQL-questionnaire, Spanish version (Ecuador) (PDQL-EV). *Parkins. Relat. Disor.* 10, 433–437. doi: 10.1016/j.parkreldis.2004.05.002

Shafazand, S., Wallace, D. M., Arheart, K. L., Vargas, S., Luca, C. C., Moore, H., et al. (2017). Insomnia, sleep quality, and quality of life in mild to moderate Parkinson's disease. *Ann. Am. Thorac. Soc.* 14, 412–419. doi: 10.1513/AnnalsATS.201608-625OC

Shalash, A. S., Hamid, E., Elrassas, H. H., Bedair, A. S., Abushouk, A. I., Khamis, M., et al. (2018). Non-motor symptoms as predictors of quality of life in Egyptian patients with Parkinson's disease: A cross-sectional study using a culturally adapted 39-item Parkinson's disease questionnaire. *Front. Neurol.* 9:357. doi: 10.3389/fneur.2018.00357

Simpson, J., Eccles, F., and Zarotti, N. (2021). *Extended evidence-based guidance on psychological interventions for psychological difficulties in individuals with Huntington's Disease*. Parkinson's Disease, Motor Neurone Disease, and Multiple Sclerosis.

Simpson, J., Lekwuwa, G., and Crawford, T. (2014). Predictors of quality of life in people with Parkinson's disease: evidence for both domain specific and general relationships. *Disabil. Rehabil.* 36, 1964–1970. doi: 10.3109/09638288.2014.883442

Sintonen, H. (1994). *The 15D-Measure of Health-Related Quality of Life. I. Reliability, Validity, and Sensitivity of Its Health State Descriptive System.* Working paper 41. Melbourne, VIC: National Centre for Health Program Evaluation. doi: 10.3109/07853890109002086

Sitzia, J., Haddrell, V., and Rice-Oxley, M. (1998). Evaluation of a nurse-led multidisciplinary neurological rehabilitation programme using the Nottingham Health Profile. *Clin. Rehabil.* 12, 389–394. doi: 10.1191/026921598675167321

Skorvanek, M., Martinez-Martin, P., Kovacs, N., Zezula, I., Rodriguez-Violante, M., Corvol, J. C., et al. (2018). Relationship between the MDS-UPDRS and quality of life: a large multicenter study of 3206 patients. *Parkinson. Relat. Disor.* 52, 83–89. doi: 10.1016/j.parkreldis.2018.03.027

Snaith, R. P., Bridge, G. W. K., and Hamilton, M. (1976). The Leeds scales for the self-assessment of anxiety and depression. *Br. J. Psychiat.* 128, 156–165. doi: 10.1192/bjp.128.2.156

Sockeel, P., Dujardin, K., Devos, D., Denève, C., Destée, A., and Defebvre, L. (2006). The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: validation in Parkinson's disease. *J. Neurol. Neurosurg. Psychiat.* 77, 579–584. doi: 10.1136/jnnp.2005.075929

Soldatos, C. R., Dikeos, D. G., and Paparrigopoulos, T. J. (2000). Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria. *J. Psychosom. Res.* 48, 555–560. doi: 10.1016/S0022-3999(00)00095-7

Soulas, T., Storme, M., Martínez-Martín, P., Pichlak, M., Gurruchaga, J. M., Palfi, S., et al. (2016). Assessing health-related quality of life with the SCOPA-PS in French individuals with Parkinson's disease having undergone DBS-STN: a validation study. *Rev. Neurol. 172,* 281–288. doi: 10.1016/j.neurol.2015.10.010

Spielberger, C., Gorsuch, R., and Lushene, R. (1970). *Manual for the State Trait Anxiety Inventory.* Palo Alto, California: Consulting Psychologist Press.

Starkstein, S. E., Mayberg, H. S., Preziosi, T. J., Andrezejewski, P., Leiguarda, R., and Robinson, R. G. (1992). Reliability, validity, and clinical correlates of apathy in Parkinson's disease. *J. Neuropsychiatry Clin. Neurosci.* 4, 134–139. doi: 10.1176/jnp.4.2.134

Starkstein, S. E., Petracca, G., Chemerinski, E., and Kremer, J. (2001). Syndromic validity of apathy in Alzheimer's disease. *Am. J. Psychiatry* 158, 872–877. doi: 10.1176/appi.ajp.158.6.872

Stiasny-Kolster, K., Mayer, G., Schäfer, S., Möller, J. C., Heinzel-Gutenbrunner, M., and Oertel, W. H. (2007). The REM sleep behavior disorder screening questionnaire–a new diagnostic instrument. *Movement Disor.* 22, 2386–2393. doi: 10.1002/mds.21740

Stocchi, F., Radicati, F. G., Chaudhuri, K. R., Johansson, A., Padmakumar, C., Falup-Pecurariu, C., et al. (2018). The Parkinson's Disease Composite Scale: results of the first validation study. *Eur. J. Neurol.* 25, 503–511. doi: 10.1111/ene.13529

Suarez, G. A., Opfer-Gehrking, T. L., Offord, K. P., Atkinson, E. J., O'Brien, P. C., and Low, P. A. (1999). The Autonomic Symptom Profile: a new instrument to assess autonomic symptoms. *Neurology* 52, 523–528. doi: 10.1212/WNL.52.3.523

Tan, L. C. S., Luo, N., Nazri, M., Li, S. C., and Thumboo, J. (2004). Validity and reliability of the PDQ-39 and the PDQ-8 in English-speaking Parkinson's disease patients in Singapore. *Parkins. Relat. Disor.* 10:493. doi: 10.1016/j.parkreldis.2004.05.007

The WHOQOL Group (1998). The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc. Sci. Med.* 46, 1569–1585. doi: 10.1016/S0277-9536(98)00009-4

Tinetti, M. E., Richman, D., and Powell, L. (1990). Falls efficacy as a measure of fear of falling. *J. Gerontol.* 45, 239–243. doi: 10.1093/geronj/45.6.P239

Tröster, A. I., Pahwa, R., Fields, J. A., Tanner, C. M., and Lyons, K. E. (2005). Quality of life in Essential Tremor Questionnaire (QUEST): development and initial validation. *Parkinson. Relat. Disor.* 11, 367–373. doi: 10.1016/j.parkreldis.2005.05.009

Tu, X. J., Hwang, W. J., Hsu, S. P., and Ma, H. I. (2017a). Responsiveness of the short-form health survey and the Parkinson's disease questionnaire in patients with Parkinson's disease. *Health Qual. Life Outc.* 15, 1–7. doi: 10.1186/s12955-017-0642-8

Tu, X. J., Hwang, W. J., Ma, H. I., Chang, L. H., and Hsu, S. P. (2017b). Determinants of generic and specific health-related quality of life in patients with Parkinson's disease. *PLoS ONE* 12:e178896. doi: 10.1371/journal.pone.0178896

Ueno, T., Kon, T., Haga, R., Nishijima, H., Arai, A., and Tomiyama, M. (2020). *Assessing the relationship between non-motor symptoms and health-related quality of life in Parkinson's* disease: a retrospective observational cohort study. *Neurol. Sci.* 41, 2867–2873. doi: 10.1007/s10072-020-04406-5

Ustün, T. B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., Epping-Jordan, J., et al. (2010). Developing the world health organization disability assessment schedule 2.0. *Bull. World Health Organ.* 88, 815–823. doi: 10.2471/BLT.09.067231

van der Eijk, M., Faber, M. J., Ummels, I., Aarts, J. W., Munneke, M., and Bloem, B. R. (2012). Patient-centeredness in PD care: development and validation of a patient experience questionnaire. *Parkinso. Relat. Disor.* 18, 1011–1016. doi: 10.1016/j.parkreldis.2012.05.017

Vandenberg, B. E., Advocat, J., Hassed, C., Hester, J., Enticott, J., and Russell, G. (2019). Mindfulness-based lifestyle programs for the self-management of Parkinson's disease in Australia. *Health Promot. Int.* 34, 668–676. doi: 10.1093/heapro/day021

Virués-Ortega, J., Carod-Artal, F. J., Serrano-Dueñas, M., Ruiz-Galeano, G., Meza-Rojas, G., Velázquez, C., et al. (2009). Cross-cultural validation of the Scales for Outcomes in Parkinson's Disease-Psychosocial questionnaire (SCOPA-PS) in four Latin American countries. *Value in Health* 12, 385–391. doi: 10.1111/j.1524-4733.2008.00436.x

Visser, M., Marinus, J., Stiggelbout, A. M., and van Hilten, J. J. (2004). Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT. *Movement Disor.* 19, 1306–1312. doi: 10.1002/mds.20153

Visser, M., van Rooden, S. M., Verbaan, D., Marinus, J., Stiggelbout, A. M., and van Hilten, J. J. (2008). A comprehensive model of health-related quality of life in Parkinson's disease. *J. Neurol.* 255, 1580–1587. doi: 10.1007/s00415-008-0994-4

Ware, J. E. Jr., Kosinski, M., and Keller, S. D. (1996). A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Med. Care* 220–233. doi: 10.1097/00005650-199603000-00003

Ware, J. E. Jr., and Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). *I. Conceptual framework and item selection. Med. Care* 30, 473–483. doi: 10.1097/00005650-199206000-00002

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54:1063. doi: 10.1037//0022-3514.54.6.1063

Webster, K., Odom, L., Peterman, A., Lent, L., and Cella, D. (1999). The functional assessment of chronic illness therapy (FACIT) measurement system: validation of version 4 of the core questionnaire. *Qual. Life Res.* 604–604.

Weintraub, D., Hoops, S., Shea, J. A., Lyons, K. E., Pahwa, R., Driver-Dunckley, E. D., et al. (2009). Validation of the questionnaire for impulsive-compulsive disorders in Parkinson's disease. *Movement Disor.* 24, 1461–1467. doi: 10.1002/mds.22571

Weitkunat, R., Letzel, H., Kanowski, S., and Grobe-Einsler, R. (1993). *Clinical and psychometric evaluation of the efficacy of nootropic drugs: characteristics of several procedures*. Zeitschrift für Gerontopsychologie und-psychiatrie.

Welsh, M., McDermott, M. P., Holloway, R. G., Plumb, S., Pfeiffer, R., Hubble, J., et al. (2003). Development and testing of the Parkinson's disease quality of life scale. *Mov. Disor.* 18, 637–645. doi: 10.1002/mds.10424

Wenning, G. K., Tison, F., Seppi, K., Sampaio, C., Diem, A., Yekhlef, F., et al. (2004). Development and validation of the Unified Multiple System Atrophy Rating Scale (UMSARS). *Mov. Disord*. 19, 1391–1402. doi: 10.1002/mds.20255

Whoqol Group (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychol. Med.* 28, 551–558. doi: 10.1017/S0033291798006667

Winter, Y., Spottke, A. E., Stamelou, M., Cabanel, N., Eggert, K., Höglinger, G. U., et al. (2011a). Health-related quality of life in multiple system atrophy and progressive supranuclear palsy. *Neurodegener. Dis.* 8, 438–446. doi: 10.1159/000325829

Winter, Y., von Campenhausen, S., Arend, M., Longo, K., Boetzel, K., Eggert, K., et al. (2011b). Health-related quality of life and its determinants in Parkinson's disease: results of an Italian cohort study. *Parkinson. Relat. Disor.* 17, 265–269. doi: 10.1016/j.parkreldis.2011.01.003

Winter, Y., von Campenhausen, S., Gasser, J., Seppi, K., Reese, J. P., Pfeiffer, K. P., et al. (2010a). Social and clinical determinants of quality of life in Parkinson's disease in Austria: a cohort study. *J. Neurol.* 257, 638–645. doi: 10.1007/s00415-009-5389-7

Winter, Y., von Campenhausen, S., Popov, G., Reese, J. P., Balzer-Geldsetzer, M., Kukshina, A., et al. (2010b). Social and clinical determinants of quality of life in Parkinson's disease in a Russian cohort study. *Parkinson. Relat. Disor.* 16, 243–248. doi: 10.1016/j.parkreldis.2009.11.012

World Health Organization (1999). *The World Health Organization Disability Assessment Schedule Phase II field Trial Instrument*. Geneva, Switzerland: World Health Organization.

Xiao, Y., Zhang, L., Wei, Q., Ou, R., Hou, Y., Liu, K., et al. (2022). Health-related quality of life in patients with multiple system atrophy using the EQ- 5D-5L. *Brain Behav*. 12:2774. doi: 10.1002/brb3.2774

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., et al. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* 17, 37–49. doi: 10.1016/0022-3956(82)90033-4

Zarotti, N., Eccles, F. J., Foley, J. A., Paget, A., Gunn, S., Leroi, I., et al. (2021). Psychological interventions for people with Parkinson's disease in the early 2020s: where do we stand? *Psychol. Psychother.* 94, 760–797. doi: 10.1111/papt.12321

Zigmond, A. S., and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta psychiatricascandinavica* 67, 361–370. doi: 10.1111/j.1600-0447.1983.tb09716.x

Zipprich, H. M., Mendorf, S., Schönenberg, A., and Prell, T. (2021). The impact of poor medication knowledge on health-related quality of life in people with Parkinson's disease: a mediation analysis. *Quality Life Res.* 1–10. doi: 10.1007/s11136-021-03024-8

Zung, W. W. (1965). A self-rating depression scale. *Arch. Gen. Psychiatry* 12, 63–70. doi: 10.1001/archpsyc.1965.01720310065008

Zung, W. W. (1972). The depression status inventory: an adjunct to the self-rating depression scale. *J. Clin. Psychol.* 28, 539–543. doi: 10.1002/1097-4679(197210)28:4<539::AID-JCLP2270280427>3.0.CO;2-S

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 September 2024
DOI 10.3389/fpsyg.2024.1395434

OPEN ACCESS

EDITED BY
Elisa Cavicchiolo,
University of Rome Tor Vergata, Italy

REVIEWED BY
Mark Ettenhofer,
University of California, San Diego, United States
Quan Wang,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Sara Cavaco
✉ sara.cavaco@chporto.min-saude.pt

RECEIVED 03 March 2024
ACCEPTED 25 July 2024
PUBLISHED 20 September 2024

CITATION
Gomes F, Ferreira I, Rosa B, Martins da Silva A and Cavaco S (2024) Using behavior and eye-fixations to detect feigned memory impairment.
Front. Psychol. 15:1395434.
doi: 10.3389/fpsyg.2024.1395434

Using behavior and eye-fixations to detect feigned memory impairment

Filomena Gomes[1,2], Inês Ferreira[2], Bruno Rosa[2], Ana Martins da Silva[3,4,5] and Sara Cavaco[1,2,3,4]*

[1]Neuropsychology Service, Centro Hospitalar Universitário de Santo António, Porto, Portugal, [2]Laboratory of Neurobiology of Human Behavior, Centro Hospitalar Universitário de Santo António, Porto, Portugal, [3]UMIB - Unit for Multidisciplinary Research in Biomedicine, ICBAS - School of Medicine and Biomedical Sciences, University of Porto, Porto, Portugal, [4]ITR - Laboratory for Integrative and Translational Research in Population Health, Porto, Portugal, [5]Department of Neurology, Centro Hospitalar Universitário de Santo António, Porto, Portugal

Background: Detecting invalid cognitive performance is an important clinical challenge in neuropsychological assessment. The aim of this study was to explore behavior and eye-fixations responses during the performance of a computerized version of the Test of Memory Malingering (TOMM-C) under standard vs. feigning conditions.

Participants and methods: TOMM-C with eye-tracking recording was performed by 60 healthy individuals (31 with standard instruction – SI; and 29 were instructed to feign memory impairment: 21 Naïve Simulators – NS and 8 Coached Simulators – CS) and 14 patients with Multiple Sclerosis (MS) and memory complaints performed. Number of correct responses, response time, number of fixations, and fixation time in old vs. new stimuli were recorded. Nonparametric tests were applied for group comparison.

Results: NS produced fewer correct responses and had longer response times in comparison to SI on all three trials. SI showed more fixations and longer fixation time on previously presented stimuli (i.e., familiarity preference) specially on Trial 1, whereas NS had more fixations and longer fixation time on new stimuli (i.e., novelty preference) specially in the Retention trial. MS patients produced longer response time and had a different fixation pattern than SI subjects. No behavioral or oculomotor difference was observed between NS and CS.

Conclusion: Healthy simulators have a distinct behavioral and eye-fixation response pattern, reflecting a novelty preference. Oculomotor measures may be useful to detect exaggeration or fabrication of cognitive dysfunction. Though, its application in clinical populations may be limited.

KEYWORDS

malingering, novelty preference, familiarity preference, eye-tracking, performance validity tests

Introduction

Malingering is the volitional feigning or exaggeration of neurocognitive symptoms for the purpose of obtaining material gain or services or avoiding formal duty, responsibility, or undesirable outcome (Slick et al., 1999; Sherman et al., 2020). The detection of noncredible cognitive performance is an important clinical challenge in neuropsychological assessment.

The presence of an external incentive (e.g., Social Security benefits, insurance compensation) is an important element to consider when distinguishing people with credible cognitive impairment from feigned cognitive deficits, even in clinical, non-forensic settings. The presence of an external incentive does not necessarily indicate unreliable neuropsychological test performance. However, it has been demonstrated that being in the process of applying for Social Security disability benefits increases the likelihood of noncredible performance (Schroeder et al., 2022; Horner et al., 2023). It has been estimated that between one third to two thirds of clinically referred patients with Social Security disability as an external incentive produce invalid data on performance validity tests - PVTs (Chafetz and Biondolillo, 2013; Schroeder et al., 2022), whereas less than one tenth of low-functioning Child Protection claimants who are motivated to do well failed on PVTs. Frequently, patients referred for routine clinical neuropsychological evaluation utilize the results of the examination for Social Security documentation.

PVTs are objective tests designed to detect invalid cognitive performance, i.e., feigned and/or exaggerated diminished capability (Sweet et al., 2021). PVTs usually require little effort or ability, as they typically are normally performed by a wide range of patients who have *bona fide* neurologic, psychiatric, or developmental problems (Heilbronner et al., 2009).

Most PVTs are forced-choice memory recognition tests and only explore accuracy to detect poor cognitive effort or malingering. Recent studies suggest that response time (Bolan et al., 2002; Kanser et al., 2019; Patrick et al., 2021b) and eye-tracking measures (Heaver and Hutton, 2011; Kanser et al., 2020; Tomer et al., 2020; Patrick et al., 2021a) may produce incremental information to the conventional accuracy responses on PVTs.

The Test of Memory Malingering (TOMM; Tombaugh, 1996) is one of the most widely used PVTs in research and clinical practice. TOMM is a forced-choice visual memory recognition test and the number of correct responses is the standard measure to discriminate between true memory impairment from noncredible performance. Response times are also able to detect feigned memory impairment on TOMM (Bolan et al., 2002; Kanser et al., 2019). The oculomotor behavior during the performance of TOMM has yet to be investigated.

This study aimed to quantify the potential information gains provided by eye fixation data in addition to behavioral response (i.e., response accuracy and response time), in the performance of a computerized version of TOMM (TOMM-C), to distinguish simulators from non-simulators. The clinical applicability of these measures was also explored in a sample of patients with multiple sclerosis and memory complaints. We hypothesized that eye-tracking metrics, in particular eye-fixations, could be an informative complement to behavioral responses on TOMM-C and that oculomotor measures are less vulnerable to coaching than behavioral responses.

Materials and methods

Subjects

Sixty healthy subjects recruited in the community were asked to perform a computerized version of TOMM (TOMM-C). The participants were distributed into two groups: 31 healthy subjects received the standard instruction (SI group) and 29 were instructed to feign memory impairment as if they were in the initial stages of dementia to benefit from Social Security disability (21 were "Naïve Simulators" – NS group, and 8 were "Coached Simulators" – CS group). Fourteen patients with diagnosis of Multiple Sclerosis (Polman et al., 2011) and with cognitive complaints on the routine neurological consultation, but without history of optic neuritis, were recruited from the outpatient clinic (MS group). All participants provided their informed written consent in accordance with the Declaration of Helsinki and the Centro Hospitalar Universitário de Santo António's Ethical Committee (reference number 026-DEFI/049-CES; Figure 1).

Procedures

TOMM-C

TOMM-C is a computerized version of the standard TOMM (Tombaugh, 1996) adapted for eye-tracking recording. TOMM-C was presented in a Windows Based Software (Presentation® - Neurobehavioral Systems, Inc.). The stimuli were presented in a TFT Monitor 19″ with touch screen (KTMT-1921-USB/B, Keytec) with behavioral response recording. The iView X™ HiSpeed 1250 System (SensoriMotoric Instruments), with chin rest and forehead rest at 45 cm from the screen, eye-movements were recorded (i.e., monocular recording) during test performance. Similar to the standard version, TOMM-C is composed of two learning trials (Trial 1 and Trial 2) and a Retention trial. In both learning trials, there was an encoding phase and a recognition phase (Figure 2). During the encoding phase, participants were shown a series of simple line drawings (i.e., the same set of stimuli as the standard TOMM; Tombaugh, 1996) for 3 s each. Between items, a cross was displayed for 1 s on the screen followed by a blank screen for 1 s. The encoding phase was immediately followed by a two-choice recognition task. A Retention Trial, which was composed by just the two-choice recognition task, was administered following a 15-min delay (without further exposure of stimulus items). During the recognition phase of the three trials, after 3 s of free viewing of each pair of drawing, participants were cued by a buzz to respond with a touch on the screen (i.e., to identify the previously seen drawing). After the subject responded, a cross was displayed for 1 s on the screen followed by a blank screen for 1 s before the display of the next item. No feedback on the accuracy of response was provided. The pattern of eye-fixations was recorded during the free viewing of the test phase. The threshold to be considered a fixation was set at 100 ms (Manor and Gordon, 2003). Two areas of interest were identified: the "old" (i.e., drawing previously seen on the learning phase) and the "new" (i.e., foil drawing). Three behavioural measures were recorded: Number of Correct Responses, Total Response Time, and Median Response Time on Correct Responses. Three oculomotor measures were considered: % of Total Number of Fixation on "new" items, % of Total Fixation Time on "new" items, and % of Fixation Time on "new" items for correct responses.

Eye-fixation data on Trial 2 and Retention Trial from two NS participants were discarded due to recording problems that resulted in extensive missing data, however their behavioral responses on those trials were analyzed. One NS participant did not produce correct responses on Trial 2, therefore the Median Response Time on Correct Responses and the % of Fixation Time on "New" for Correct Responses could not be calculated.

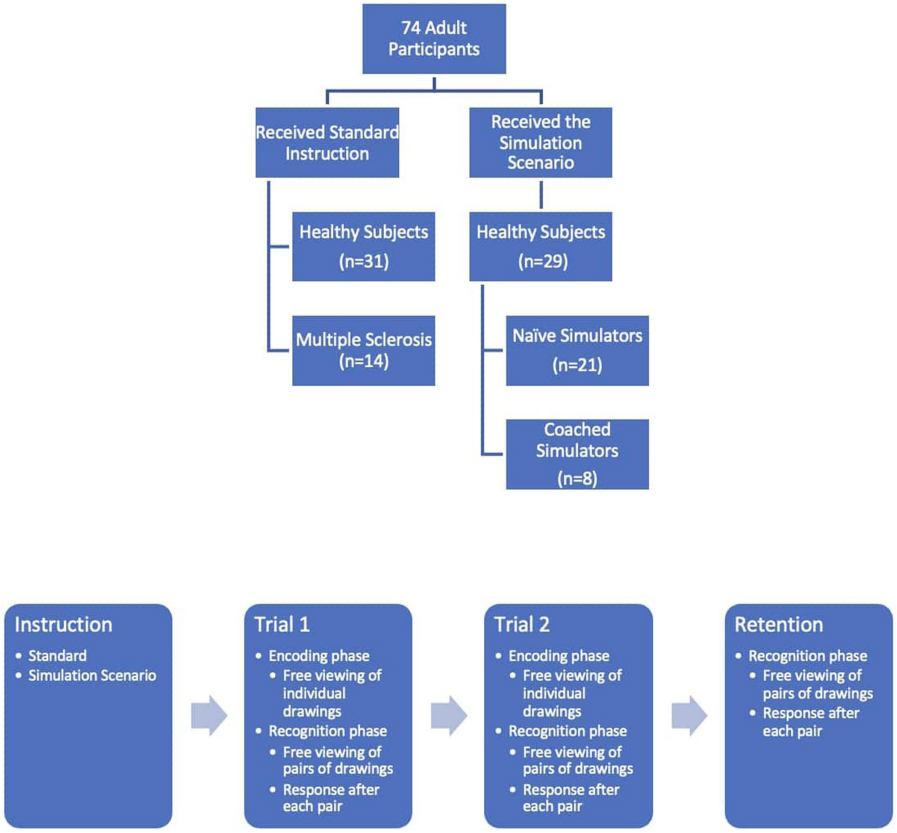


FIGURE 1
Flowchart of the study participants and of the TOMM experiment.

Standard instruction and multiple sclerosis

SI and MS participants were asked to perform the TOMM-C to the best of their ability. The MS group were also asked to perform the Nine Hole Peg Test, the Symbol Digit Modalities Test – SDMT and the Auditory Verbal Learning Test - AVLT. The SDMT (Sousa et al., 2021) and AVLT (Cavaco et al., 2015) were adjusted to the demographic characteristics of the subjects according to the available norms.

Naïve and coached simulators

Both NS and CS participants were read the a scenario in which they were asked to imagine experiencing real memory difficulties and feeling no longer competent to carry out their work; and to request disability benefits they would need to go through a neuropsychological assessment and convince the examiner of their disability by highlighting their memory difficulties in a credible way. Following the literature (Frederick and Foster, 1991; Rüsseler et al., 2008; Jones, 2017), the CS participants additionally received a series of suggestions to produce the most severe memory problems *without* making it too obvious to the examiner.

Statistical analyses

Descriptive statistics and nonparametric tests (Chi-square test or Fisher's exact test and Mann–Whitney test) were used for characterization and comparison of the groups. Effect sizes were calculated and interpreted as follows: 0.2 (small), 0.5 (medium), and 0.8 (large). Receiver operating characteristic (ROC) curves were applied to differentiate SI and NS participants on each measure and trial. The area under the curve was calculated. By design, PVTs prioritize specificity over sensitivity and it is recommended that PVTs have at least 90% specificity when applied to individuals with evidence of significant cognitive dysfunction (Sherman et al., 2020; Sweet et al., 2021). Therefore, the specificity was set at ≥90%. The sensitivity, positive predictive value and negative predictive value were calculated. The cut-off scores were then used to identify frequency of abnormal score in MS and CS groups.

Multiple logistic regression analyses were used to explore the association between abnormal TOMM score performance and group, while taking into consideration demographic characteristics. TOMM score (recoded according to the cut-off) was the dependent variable, whereas group (i.e., SI vs. MS), sex, age, and years of education were the independent variables. No variable selection was applied and basic assumptions were verified. Simple logistic regression analyses were used to explore in MS group the association between performance on the SDMT and AVLT and some TOMM-C measures.

Results

Groups characteristics

As presented in Table 1, the SI group ($n = 31$) and the NS group ($n = 21$) had similar demographic characteristics, namely sex, age, and

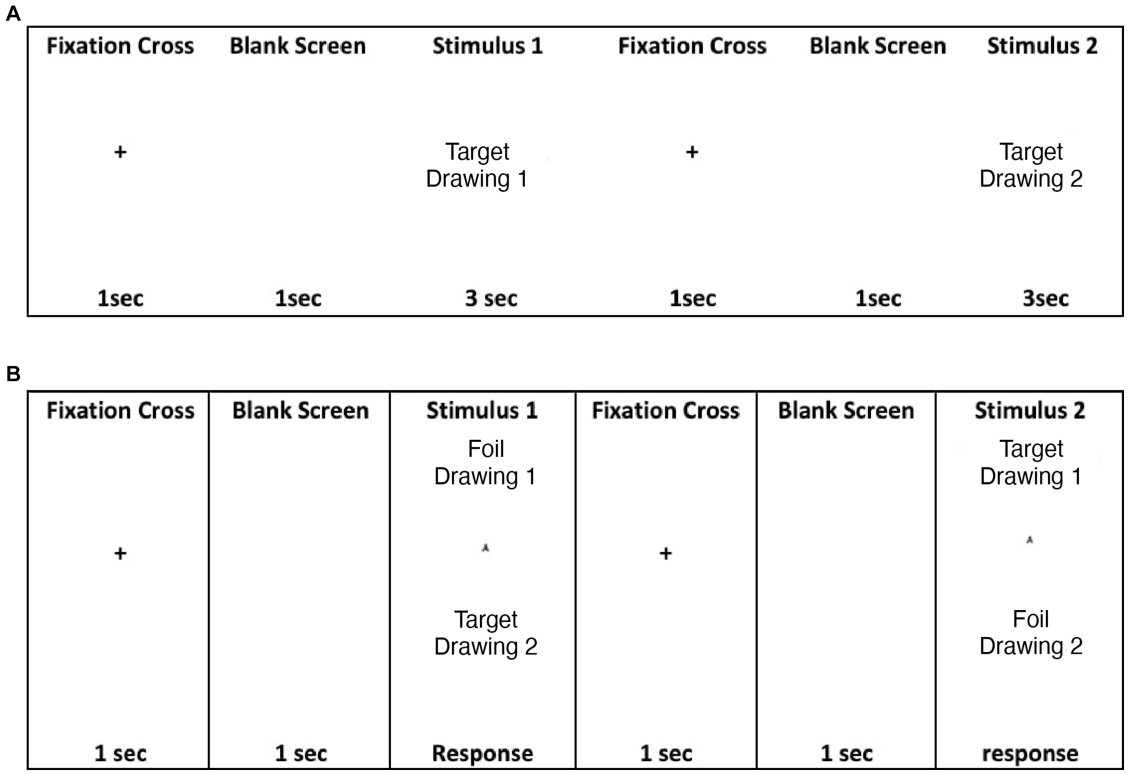


FIGURE 2
(A) TOMM encoding phase at Trials 1 and 2; **(B)** TOMM two choice recognition phase at Trial 1, Trial 2, and Retention.

education. SI group was younger than the MS group ($n = 14$; $p = 0.018$) and had fewer years of education than the CS group ($n = 8$; $p = 0.031$). NS individuals were younger ($p < 0.001$) and had more years of education ($p = 0.011$) than MS patients; and were younger than CS ($p = 0.029$) participants. No group differences were recorded regarding sex.

MS patients median *T*-score (25th, 75th percentiles) on the SDMT was 44.1 (31.0, 50.5). The median adjusted score (25th, 75th percentiles) of the AVLT-Delayed Recall was −1.0 (−1.7, 0.2). These adjusted scores correspond to the number of standard deviations below/above the mean of the participant's normal peers with the same sex, age, and education. Three MS patients (21.4%) scored below the estimated 18th percentile on AVLT-Delayed Recognition for the demographic characteristics of each individual.

TOMM-C performance

As shown in Table 1, the NS group produced fewer correct responses (large effect sizes), had longer total response time and median response time on correct responses (small effect sizes), higher % of total number of fixations on "new" and % of total fixation time on "new" stimuli (medium effect sizes for Trials 1 and Retention, and small effects for Trial 2), and % of fixation time on "new" stimuli for correct responses (small effect sizes on all trials) in comparison to the SI group on all TOMM-C trials.

In comparison to the SI group, MS participants were slower to respond (i.e., total response time and median response time for correct responses) on all trials ($p < 0.001$), but had similar number of correct responses. The effect sizes for the response time measures were relatively small. On both Trial 1 ($p < 0.06$) and Retention Trial ($p < 0.01$), the % of total number of fixations, the % of total fixation time on "new," and the % of fixation time on "new" for correct responses was higher for the MS group than the SI group, with a small effect size. The MS group only differed from the NS group on the number of correct responses (all $p < 0.001$ and with large effect sizes). Neither response time nor oculomotor measures differed between MS and NS participants.

NS and CS groups did not differ on any of TOMM-C behavioral and oculomotor measures. As shown in Table 1, the comparison between SI group and CS group revealed significant differences on all measures, but only partially (i.e., not on all trials). Regarding the number of correct responses, a medium effect size was observed for Trial 1 and large effect sizes were recorded for Trials 2 and Retention. For both Total Response Time and Median Response Time for Correct Responses, the effect sizes were relatively small. Medium effect sizes were observed for Trial 1 on the % of Total Number of Fixation on "New" and % of Total Fixation Time on "New." Small effect sizes were also recorded on these measures at Retention Trial.

Table 2 shows the best cut-off scores to differentiate SI and NS participants, while setting the specificity at >90%. These cut-off scores were then used to identify the frequency of abnormal scores in the MS group and CS group.

Multiple logistic regression analyses revealed that, while taking into consideration demographic characteristics, MS patients had higher odds of abnormal score than SI participants on Total Response

TABLE 1 Demographic characterization and TOMM-C behavioral and eye-fixation scores.

			Standard instruction (SI) group (n = 31)	Naive simulators (NS) group (n = 21)	SI vs. NS		Multiple Sclerosis (MS) group (n = 14)	SI vs. MS		NS vs. MS		Coached simulators (CS) group (n = 8)	SI vs. CS		NS vs. CS	
					p	Effect size		p	Effect size	p	Effect size		p	Effect size	p	Effect size
Sex	Female		27 (87.1%)	19 (90.5%)	>0.999	0.05	11 (78.6%)	0.659	0.11	0.369	0.17	6 (75.0%)	0.583	0.14	0.300	0.20
Age	Years		31 (28, 46)	30 (28, 33)	0.182	0.19	45 (37, 52)	0.018	0.35	<0.001	0.67	35 (30, 45)	0.465	0.12	0.029	0.40
Education	Years		16 (15, 18)	17 (16, 18)	0.185	0.18	15 (12, 17)	0.185	0.20	0.011	0.43	18 (17, 21)	0.031	0.35	0.077	0.33
TOMM-C	Correct responses	Trial 1	49 (46, 50)	25 (18, 33)	<0.001	0.84	49 (47, 50)	0.940	0.01	<0.001	0.84	27 (20, 29)	<0.001	0.70	0.941	0.01
		Trial 2	50 (50, 50)	28 (22, 36)	<0.001	0.94	50 (50, 50)	0.502	0.10	<0.001	0.86	33 (30, 38)	<0.001	0.94	0.171	0.25
		Retention trial	50 (50, 50)	26 (17, 34)	<0.001	0.94	50 (50, 50)	0.559	0.09	<0.001	0.86	25 (17, 28)	<0.001	0.94	0.732	0.06
	Total response time	Trial 1	80.6 (69.7, 103.4)	94.3 (82.1, 134.5)	0.031	0.30	110.5 (86.8, 130.2)	0.005	0.42	0.711	0.06	102.5 (95.9, 119.1)	0.020	0.37	0.591	0.02
		Trial 2	66.5 (56.7, 83.6)	97.6 (64.7, 132.6)	0.009	0.36	94.3 (76.8, 112.9)	0.005	0.42	0.893	0.02	88.7 (75.9, 114.0)	0.018	0.38	0.770	0.05
		Retention trial	59.9 (46.5, 80.1)	88.6 (66.8, 105.6)	<0.001	0.48	88.4 (64.8, 101.8)	0.003	0.44	0.736	0.06	84.9 (67.4, 101.6)	0.044	0.32	0.845	0.04
	Median response time for correct responses	Trial 1	1.5 (1.2, 1.8)	1.8 (1.4, 2.4)	0.031	0.30	1.9 (1.4. 2.2)	0.012	0.38	0.920	0.02	1.9 (1.6, 2.3)	0.014	0.40	0.807	0.05
		Trial 2	1.3 (1.0, 1.6)	1.8 (1.2, 2.6)	0.007	0.37	1.7 (1.5, 1.9)	0.008	0.39	0.576	0.09	1.6 (1.5, 1.9)	0.031	0.34	0.684	0.08
		Retention trial	1.1 (0.9, 1.5)	1.6 (1.3, 2.0)	<0.001	0.48	1.7 (1.1, 1.9)	0.009	0.39	0.662	0.07	1.7 (1.2, 2.0)	0.037	0.33	0.884	0.03
	% of Total number of fixation on "New"	Trial 1	43.9 (39.1, 49.6)	50.7 (49.1, 53.2)	<0.001	0.66	49.2 (42.2, 52.4)	0.056	0.29	0.114	0.27	51.7 (50.0, 54.3)	<0.001	0.59	0.591	0.10
		Trial 2	47.0 (42.4, 52.9)	52.2 (50.6, 55.3)	0.016	0.34	49.1 (48.3, 54.0)	0.135	0.22	0.248	0.20	51.5 (43.3, 54.0)	0.554	0.09	0.387	0.16
		Retention trial	44.9 (41.1, 54.0)	53.8 (48.5, 59.8)	<0.001	0.51	54.1 (47.1, 59.1)	0.008	0.40	0.576	0.10	54.6 (51.3, 59.0)	0.004	0.46	0.799	0.15
	% of Total fixation time on "New"	Trial 1	43.1 (38.1, 47.4)	51.3 (49.6, 53.1)	<0.001	0.70	48.1 (41.5, 52.2)	0.047	0.30	0.055	0.33	51.1 (48.2, 52.6)	<0.001	0.53	0.696	0.07
		Trial 2	45.7 (39.4, 53.2)	53.6 (49.8, 55.4)	0.009	0.36	50.6 (47.3, 55.6)	0.148	0.22	0.401	0.14	51.0 (40.6, 55.9)	0.531	0.10	0.309	0.19
		Retention trial	44.1 (38.6. 52.7)	55.3 (48.0, 59.7)	<0.001	0.53	53.4 (46.5, 59.7)	0.007	0.42	0.552	0.10	54.8 (51.4. 61.1)	0.003	0.47	0.760	0.06
	% of Fixation time on "New" for correct responses	Trial 1	42.3 (38.0, 46.8)	50.2 (45.4, 53.2)	<0.001	0.46	48.1 (41.3, 51.8)	0.050	0.29	0.312	0.17	47.7 (40.1, 53.0)	0.082	0.28	0.464	0.14
		Trial 2	45.7 (39.4, 53.2)	51.3 (47.3, 57.4)	0.084	0.24	50.6. (47.3, 55.6)	0.148	0.22	0.942	0.01	47.2 (31.0, 55.2)	0.972	0.01	0.367	0.17
		Retention trial	44.1 (38.6, 52.7)	52.4 (46.5, 62.0)	0.007	0.38	53.4 (46.5, 60.0)	0.007	0.40	0.916	0.02	54.7 (47.9, 57.7)	0.047	0.32	0.879	0.03

Group comparisons. Data are presented as frequencies (%) and medians (25th, 75th percentiles). Chi-square test (or Fisher's exact), Mann–Whitney test, and effect sizes were applied for group comparison. One Naïve Simulator did not produce correct responses on Trial 2. Missing data: eye-fixation data of Trial 2 and Retention Trial of two Naïve Simulator.

10.3389/fpsyg.2024.1395434

TABLE 2 Diagnostic statistics of TOMM-C behavioral and eye-fixation measures, in the comparison between healthy with standard Instruction (SI) and healthy with naïve simulation instruction (NS) groups.

			Heathy with standard instruction vs. Healthy with naïve simulation instruction								Frequency of abnormal score	
			AUC	95% CI	*p*	Cut-off Point	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Multiple sclerosis with standard instruction (MS) group (*n* = 14)	Healthy with coached simulation instruction (*n* = 8)
TOMM-C	Correct responses	Trial 1	0.993	0.979, 1.000	<0.001	45	93.5%	100%	91.3%	100%	0 (0%)	8 (100%)
		Trial 2	1.000	1.000, 1.000	<0.001	49	100%	100%	100%	100%	0 (0%)	8 (100%)
		Retention trial	1.000	1.000, 1.000	<0.001	49	100%	100%	100%	100%	0 (0%)	8 (100%)
	total response time	Trial 1	0.677	0.523, 0.832	0.031	122.5	33.3%	90.3%	70.0%	66.7%	5 (35.7%)	1 (12.5%)
		Trial 2	0.716	0.566, 0.865	0.009	97.0	52.4%	90.3%	78.6%	73.7%	7 (50.0%)	2 (25.0%)
		Retention trial	0.786	0.658, 0.915	0.001	100.6	33.3%	93.5%	77.8%	67.4%	4 (28.6%)	2 (25.0%)
	Median response time for correct responses	Trial 1	0.677	0.517, 0.838	0.031	2.1	28.6%	93.5%	75.0%	65.9%	5 (35.7%)	3 (37.5%)
		Trial 2	0.724	0.577, 0.872	0.007	1.8	50.0%	90.3%	76.9%	73.7%	5 (35.7%)	3 (37.5%)
		Retention trial	0.785	0.658, 0.912	0.001	2.0	23.8%	93.5%	71.4%	64.4%	2 (14.3%)	1 (12.5%)
	% of Total number of fixation on "New"	Trial 1	0.890	0.805, 0.975	<0.001	50.1	61.9%	96.8%	92.9%	78.9%	6 (42.9%)	6 (75.0%)
		Trial 2	0.739	0.602, 0.875	0.005	55.8	15.8%	90.3%	50.0%	63.6%	3 (21.4%)	0 (0%)
		Retention TRIAL	0.806	0.688, 0.925	<0.001	57.8	26.3%	90.3%	62.5%	66.7%	4 (28.6%)	2 (25.0%)
	% of Total fixation time on "New"	Trial 1	0.916	0.836, 0.995	<0.001	49.2	81.0%	93.5%	89.5%	87.9%	7 (50.0%)	6 (75.0%)
		Trial 2	0.756	0.623, 0.888	0.003	58.8	10.5%	90.3%	40.0%	62.2%	0 (0.0%)	1 (12.5%)
		Retention trial	0.818	0.704, 0.933	<0.001	57.8	35.0%	93.5%	77.8%	70.7%	4 (28.6%)	3 (37.5%)
	% of Fixation time on "New" for correct responses	Trial 1	0.774	0.636, 0.913	0.001	49.7	52.4%	93.5%	84.6%	74.4%	6 (42.9%)	3 (37.5%)
		Trial 2	0.683	0.534, 0.832	0.034	58.6	16.7%	90.3%	50.0%	65.1%	0 (0.0%)	1 (12.5%)
		Retention trial	0.730	0.580, 0.880	0.007	57.8	26.3%	93.5%	71.4%	67.4%	4 (28.6%)	2 (25.0%)

Frequency of abnormal score for the multiple sclerosis with standard instruction (MS) group and the healthy with coached simulation (CS) groups. AUC, the area under de curve. For all measures, except correct responses, scores > the cut-off point were considered abnormal. Missing data: Eye-fixation data on Trial 2 and Retention Trial from two Naïve Simulators were discarded due to recording problems that resulted in extensive missing data. One Naïve Simulator did not produce correct responses on Trial 2, therefore the Median Response Time on Correct Responses and the % of Fixation Time on "New" for Correct Responses could not be calculated.

10.3389/fpsyg.2024.1395434

Time and Median Response Time for Correct Responses at Trial 1 (respectively adjusted odds = 6.441, p = 0.076 and adjusted odds = 25.027, p = 0.016) and Trial 2 (respectively adjusted odds = 11.001, p = 0.008 and adjusted odds = 4.476, p = 0.086). MS patients also had higher odds of abnormal score at Trial 1 on the following oculomotor measures: % of Total Number of Fixation on "New" (adjusted odds = 44.085, p = 0.005), % of Total Fixation Time on "New" (adjusted odds = 34.961, p = 0.003), and % of Fixation Time on "New" for Correct Responses (adjusted odds = 40.412, p = 0.007). No statistically significant difference (p > 0.05) on oculomotor measures was found between SI and MS participants on Trial 2 and Retention trial, when demographic characteristics were considered.

Simple logistic regressions were used to explore in MS patients the association between standard neuropsychological measures (i.e., SDMT and AVLT) and the following TOMM-C measures: Total Response Time, Median Response Time for Correct Responses, and % of Total Fixation Time on "New." No significant association was found (p > 0.05).

Discussion

Study results revealed that both behavioral responses (i.e., response accuracy and response time) and eye-fixation data can distinguish simulators from non-simulators in a computerized version of TOMM. Healthy simulators were asked to imagine experiencing "real" memory problems and needing to exaggerate their cognitive difficulties to obtain disability benefits.

Eye-fixation recordings of the SI group showed a familiarity preference (i.e., shorter fixation time on "new" stimuli), especially on Trial 1, whereas both simulator groups showed a novelty preference (i.e., longer fixations on "new" stimuli than on previously presented stimuli) on the three TOMM-C trials. The eye-fixation measure with the best diagnostic statistics in differentiating SI from NS participants was % of Total Fixation Time on "New" (sensitivity of 81.0% and specificity of 93.5%). These findings are consistent with a recent study (Tomer et al., 2020), which revealed that simulators spent more time gazing at foils than target stimuli in another PVT - the Word Memory Test. In non-clinical samples, a novelty preference appears to be a marker of non-credible performance on PVTs that require forced-choice recognition. It has also been suggested that visual disengagement (i.e., gaze aversion) may be used by simulators to attenuate visual input and thereby decrease the cognitive load that they may be experiencing while performing the test (Tomer et al., 2020). Though, gaze aversion could not be documented in the present study, because only two areas of interest - AOI (i.e., the screen was divided in two - "old"/ "new" drawings) were considered, fixations in non-relevant spaces within each AOI were considered on target, and fixations outside the two AOI were discarded. Future studies ought to explore in greater detail the viewing pattern during the performance of TOMM.

Forced-choice memory recognition PVTs (e.g., TOMM and Word Memory Test) share some resemblance with visual-paired comparison (VPC) tasks, which were designed to measure infant recognition memory. Both typically involve a familiarization phase followed by a test phase. During the familiarization phase, the individual is presented with a set of visual stimuli. During the test phase, the familiarization stimulus is paired with a novel stimulus. On VPC tasks, the spontaneous eye-movements are recorded and the amount of time spent looking at each stimulus during the test phase is usually the primary dependent variable. A decreased attention to familiar patterns relative to novel ones (i.e., spending more time looking at novel images) has been observed in VPC applied to preverbal human infants (Fantz, 1964), human adults (Manns et al., 2000), and primates (Pascalis and Bachevalier, 1999). VPC may also elicit a preference for familiarity depending on the length of the retention interval (Bahrick and Pickens, 1995; Richmond et al., 2007). Unlike standard VPC tasks, forced-choice memory recognition tests require an explicit recognition instruction and the visual behaviour of healthy adults during the test phase has been shown to favour familiar stimuli (Richmond et al., 2007; Brooks et al., 2023). Both the preference for novelty and the preference for familiarity are usually interpreted as evidence of recognition memory, whereas null preferences can be interpreted as evidence of forgetting (Richmond et al., 2007).

MS patients exhibited a less evident familiarity preference on the eye-fixation data than the SI participants on Trial 1. It's unclear why half of the patients with MS showed a preference for the "New" stimuli, as measured by the % of Total Fixation Time on "New." Both the preference for novelty and the preference for familiarity are usually interpreted as evidence of recognition memory, whereas null preferences can be interpreted as evidence of forgetting (Richmond et al., 2007). It is reasonable to speculate that MS patients were more alert to the possibility that novel stimuli might be relevant, because of their prior experience with neuropsychological assessments (for clinical purposes) that require recall and recognition of previously presented stimuli without prior warning. In the MS group, the % of Total Fixation Time on "New" was not related to measures of visual working memory/ psychomotor speed (i.e., SDMT) and verbal memory (i.e., AVLT Delayed Recall and Delayed Recognition), even though patients' performance on these standard neuropsychological measures was as expected mildly below the norm (Martins Da Silva et al., 2015). Future studies ought to explore the preference for familiarity / novelty in *bona fide* MS patients and in other clinical populations and to investigate their associations with standard measures of memory (both visual and verbal).

The number of Correct Responses produced the most robust diagnostic statistics and the identified cut-off scores are consistent with most studies in the literature that explored simulation in healthy individuals (for a review see: Martin et al., 2020). Healthy individuals feigning memory impairment significantly produced fewer correct responses on TOMM-C than the SI group. NS and CS performance on TOMM-C approached chance level, especially on Trial 1 and Retention Trial. A ceiling effect was observed in healthy individuals with credible performance (Rees et al., 1998). The number of Correct Responses was similar between MS patients and SI healthy individuals on all trials. Furthermore, only the number of Correct Responses clearly differentiated MS patients from NS participants. These results provide support to its use in clinical practice, namely in patients with MS, cognitive complaints, and mild memory difficulties.

Response time differentiated SI participants from both simulator groups, confirming previous reports (Bolan et al., 2002; Kanser et al., 2019) that healthy simulators are slower to respond on TOMM. However, MS patients were also slower to respond than healthy individuals under SI condition and had similar latency to the NS group. These results highlight the need for caution when applying response time as a performance validity measure in clinical populations, namely in MS which is known to produce processing speed deficits in most patients (Ruano et al., 2017). Nonetheless, in the present study no clear association was found between response time on TOMM-C and a standard measure of visual working memory and psychomotor speed (i.e., SDMT).

No effect of coaching how to avoid detection of invalid performance was observed on any of the behavioural and eye-fixation measures of TOMM-C. In other words, the performance of NS and CS participants did not differ, which may reflect lack of statistical power or resistance of the test to coaching (Jelicic et al., 2011). Larger samples are necessary to confirm these negative findings.

The simultaneous recording of both behavioral and eye-fixation measures in one of the most widely used PVTs, the exploration of different experimental conditions, and the inclusion of a clinical sample are strengths of the study. Unfortunately, the inclusion of participants was cut short due to equipment failure. As ensuing, the small size of the studied groups and the demographic differences of the groups (namely regarding age and education) limit the informative value of group comparisons and the generalizability of the research findings. Though, the literature has recorded minimal or no effects of age or education on TOMM performance (Rees et al., 1998; Rai and Erdodi, 2021; Tchienga and Golden, 2022). Furthermore, the characteristics of the clinical group were not ideal, because none of the MS participants with cognitive complaints had a diagnosis of dementia and not all had memory impairment. Future studies ought to study other clinical aetiologies and suspected clinical malingers.

Recent studies with pupillometry have reported that pupil dilation can detect feigned cognitive impairment on TOMM (Heaver and Hutton, 2011; Patrick et al., 2021a,b). However, the present study focused only on eye-fixations. Future studies should explore the possibility of combining pupil reactivity with eye-fixation pattern in the detection of deception. The standardization of the viewing period (3 s) prior to the behavioral response facilitated the comparison between participants, though it also limited the informative value of the response time.

In sum, healthy individuals feigning memory impairment showed a distinct behavioral (i.e., fewer correct responses and longer response times) and oculomotor (i.e., longer fixation time on "new" stimuli) response pattern on a computerized version of TOMM, which may reflect an increased effort to inhibit a natural response. Further investigation is necessary to understand the potential application of response time and eye-fixation measures in real-life clinical situations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Comissão de Ética do Centro Hospitalar Universitário de Santo António. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

FG: Writing – original draft, Writing – review & editing, Investigation. IF: Investigation, Writing – review & editing, Writing – original draft. BR: Software, Writing – review & editing, Writing – original draft. AM: Methodology, Writing – review & editing, Writing – original draft. SC: Methodology, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by Bial Foundation Grant 430/14 UIDB/00215/2020; UIDP/00215/2020; LA/P/0064/2020.

Acknowledgments

We would like to express our gratitude for Ana Filipa Gerós and Paulo de Castro Aguiar contribution to data processing.

Conflict of interest

The 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.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1395434/full#supplementary-material

References

Bahrick, L. E., and Pickens, J. N. (1995). Infant memory for object motion across a period of three months: implications for a four-phase attention function. *J. Exp. Child Psychol.* 59, 343–371. doi: 10.1006/jecp.1995.1017

Bolan, B., Foster, J. K., Schmand, B., and Bolan, S. (2002). A comparison of three tests to detect feigned amnesia: the effects of feedback and the measurement of response latency. *J. Clin. Exp. Neuropsychol.* 24, 154–167. doi: 10.1076/jcen.24.2.154.1000

Brooks, G., Whitehead, H., and Köhler, S. (2023). When familiarity not novelty motivates information-seeking behaviour. *Sci. Rep.* 13:5201. doi: 10.1038/s41598-023-31953-6

Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., et al. (2015). Auditory verbal learning test in a large nonclinical Portuguese population. *Appl. Neuropsychol. Adult* 22, 321–331. doi: 10.1080/23279095.2014.927767

Chafetz, M. D., and Biondolillo, A. M. (2013). Feigning a severe impairment profile. *Arch. Clin. Neuropsychol.* 28, 205–212. doi: 10.1093/arclin/act015

Fantz, R. L. (1964). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science* 146, 668–670. doi: 10.1126/science.146.3644.668

Frederick, R. I., and Foster, H. G. (1991). Multiple measures of malingering on a forced-choice test of cognitive ability. *Psychol. Assess. J. Consult. Clin. Psychol.* 3, 596–602. doi: 10.1037/1040-3590.3.4.596

Heaver, B., and Hutton, S. B. (2011). Keeping an eye on the truth? Pupil size changes associated with recognition memory. *Memory* 19, 398–405. doi: 10.1080/09658211.2011.575788

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., and Millis, S. R.Conference Participants (2009). American academy of clinical neuropsychology consensus conference statement on the neuropsychological assessment of effort, response Bias, and malingering. *Clin. Neuropsychol.* 23, 1093–1129. doi: 10.1080/13854040903155063

Horner, M. D., Denning, J. H., and Cool, D. L. (2023). Self-reported disability-seeking predicts PVT failure in veterans undergoing clinical neuropsychological evaluation. *Clin. Neuropsychol.* 37, 387–401. doi: 10.1080/13854046.2022.2056923

Jelicic, M., Ceunen, E., Peters, M. J. V., and Merckelbach, H. (2011). Detecting coached feigning using the test of memory malingering (TOMM) and the structured inventory of malingered symptomatology (SIMS). *J. Clin. Psychol.* 67, 850–855. doi: 10.1002/jclp.20805

Jones, S. M. (2017). Dissimulation strategies on standard neuropsychological tests: a qualitative investigation. *Brain Inj.* 31, 1131–1141. doi: 10.1080/02699052.2017.1283444

Kanser, R. J., Bashem, J. R., Patrick, S. D., Hanks, R. A., and Rapport, L. J. (2020). Detecting feigned traumatic brain injury with eye tracking during a test of performance validity. *Neuropsychology* 34, 308–320. doi: 10.1037/neu0000613

Kanser, R. J., Rapport, L. J., Bashem, J. R., and Hanks, R. A. (2019). Detecting malingering in traumatic brain injury: combining response time with performance validity test accuracy. *Clin. Neuropsychol.* 33, 90–107. doi: 10.1080/13854046.2018.1440006

Manns, J. R., Stark, C. E. L., and Squire, L. R. (2000). The visual paired-comparison task as a measure of declarative memory. *Proc. Natl. Acad. Sci.* 97, 12375–12379. doi: 10.1073/pnas.220398097

Manor, B. R., and Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods* 128, 85–93. doi: 10.1016/S0165-0270(03)00151-1

Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N., et al. (2020). A systematic review and meta-analysis of the test of memory malingering in adults: two decades of deception detection. *Clin. Neuropsychol.* 34, 88–119. doi: 10.1080/13854046.2019.1637027

Martins Da Silva, A., Cavaco, S., Moreira, I., Bettencourt, A., Santos, E., Pinto, C., et al. (2015). Cognitive reserve in multiple sclerosis: protective effects of education. *Mult. Scler. J.* 21, 1312–1321. doi: 10.1177/1352458515581874

Pascalis, O., and Bachevalier, J. (1999). Neonatal aspiration lesions of the hippocampal formation impair visual recognition memory when assessed by paired-comparison task but not by delayed nonmatching-to-sample task. *Hippocampus* 9, 609–616. doi: 10.1002/(SICI)1098-1063(1999)9:6<609::AID-HIPO1>3.0.CO;2-A

Patrick, S. D., Rapport, L. J., Kanser, R. J., Hanks, R. A., and Bashem, J. R. (2021a). Detecting simulated versus bona fide traumatic brain injury using pupillometry. *Neuropsychology* 35, 472–485. doi: 10.1037/neu0000747

Patrick, S. D., Rapport, L. J., Kanser, R. J., Hanks, R. A., and Bashem, J. R. (2021b). Performance validity assessment using response time on the Warrington recognition memory test. *Clin. Neuropsychol.* 35, 1154–1173. doi: 10.1080/13854046.2020.1716997

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* 69, 292–302. doi: 10.1002/ana.22366

Rai, J. K., and Erdodi, L. A. (2021). Impact of criterion measures on the classification accuracy of TOMM-1. *Appl. Neuropsychol. Adult* 28, 185–196. doi: 10.1080/23279095.2019.1613994

Rees, L. M., Tombaugh, T. N., Gansler, D. A., and Moczynski, N. P. (1998). Five validation experiments of the test of memory malingering (TOMM). *Psychol. Assess.* 10, 10–20. doi: 10.1037/1040-3590.10.1.10

Richmond, J., Colombo, M., and Hayne, H. (2007). Interpreting visual preferences in the visual paired-comparison task. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 823–831. doi: 10.1037/0278-7393.33.5.823

Ruano, L., Portaccio, E., Goretti, B., Niccolai, C., Severo, M., Patti, F., et al. (2017). Age and disability drive cognitive impairment in multiple sclerosis across disease subtypes. *Mult. Scler. J.* 23, 1258–1267. doi: 10.1177/1352458516674367

Rüsseler, J., Brett, A., Klaue, U., Sailer, M., and Münte, T. F. (2008). The effect of coaching on the simulated malingering of memory impairment. *BMC Neurol.* 8:37. doi: 10.1186/1471-2377-8-37

Schroeder, R. W., Clark, H. A., and Martin, P. K. (2022). Base rates of invalidity when patients undergoing routine clinical evaluations have social security disability as an external incentive. *Clin. Neuropsychol.* 36, 1902–1914. doi: 10.1080/13854046.2021.1895322

Sherman, E. M. S., Slick, D. J., and Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: a 20-year update of the malingered neuropsychological dysfunction criteria. *Arch. Clin. Neuropsychol.* 35, 735–764. doi: 10.1093/arclin/acaa019

Slick, D. J., Sherman, E. M. S., and Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: proposed standards for clinical practice and research. *Clin. Neuropsychol.* 13, 545–561. doi: 10.1076/1385-4046(199911)13:04;1-Y;FT545

Sousa, C., Rigueiro-Neves, M., Passos, A. M., Ferreira, A., and Sá, M. J.Group for Validation of the BRBN-T in the Portuguese MS Population (2021). Assessment of cognitive functions in patients with multiple sclerosis applying the normative values of the Rao's brief repeatable battery in the Portuguese population. *BMC Neurol.* 21:170. doi: 10.1186/s12883-021-02193-w

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., et al. (2021). American Academy of clinical neuropsychology (AACN) 2021 consensus statement on validity assessment: update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *Clin. Neuropsychol.* 35, 1053–1106. doi: 10.1080/13854046.2021.1896036

Tchienga, I., and Golden, C. (2022). A-235 the degree to which age, education and race predict TOMM performance in a retired NFL cohort. *Arch. Clin. Neuropsychol.* 37:1391. doi: 10.1093/arclin/acac060.235

Tombaugh, T. N. (1996). Test of memory malingering (TOMM). North Tonawanda, NY: Multi Health Systems.

Tomer, E., Lupu, T., Golan, L., Wagner, M., and Braw, Y. (2020). Eye tracking as a mean to detect feigned cognitive impairment in the word memory test. *Appl. Neuropsychol. Adult* 27, 49–61. doi: 10.1080/23279095.2018.1480483

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 23 September 2024
DOI 10.3389/fpsyg.2024.1399388

OPEN ACCESS

EDITED BY
Elisa Cavicchiolo,
University of Rome Tor Vergata, Italy

REVIEWED BY
James Hugo Smith-Spark,
London South Bank University,
United Kingdom
Srishti Nayak,
Vanderbilt University Medical Center,
United States

*CORRESPONDENCE
Vinícius Figueiredo de Oliveira
✉ viniciusfo96@gmail.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 11 March 2024
ACCEPTED 25 July 2024
PUBLISHED 23 September 2024

CITATION
de Oliveira VF, Vial-Martins J, Pinto ALCB, Fonseca RP and Malloy-Diniz LF (2024) A new neuropsychological tool for simultaneous reading and executive functions assessment: initial psychometric properties.
Front. Psychol. 15:1399388.
doi: 10.3389/fpsyg.2024.1399388

A new neuropsychological tool for simultaneous reading and executive functions assessment: initial psychometric properties

Vinícius Figueiredo de Oliveira[1*†], Jéssica Vial-Martins[1†], André Luiz de Carvalho Braule Pinto[2], Rochele Paz Fonseca[3] and Leandro Fernandes Malloy-Diniz[1]

[1]Laboratory of Medical Psychology and Neuropsychology, Department of Mental Health, Faculty of Medicine, Federal University of Minas Gerais, Belo Horizonte, Brazil, [2]Amazonas Psychological Assessment Laboratory, Faculty of Psychology, Federal University of Amazonas, Manaus, Brazil, [3]Experimental and School Neuropsychology, Faculty of Medicine, Federal University of Minas Gerais, Belo Horizonte, Brazil

Introduction: The development of reading and complex executive functions is fundamental for achieving social, academic, and professional success. So far, there is no single neuropsychological instrument that comprehensively assesses the domains of inhibitory control, cognitive flexibility, working memory, and reading comprehension. To assess executive functions related to reading, the "Assessment of Reading and Executive Functions" (AREF) was developed. In this study, we show initial evidence of validity and reliability for four subtests - Graphophonological-Semantic Flexibility, Inhibitory Control, Flexibility, and Working Memory.

Methods: A total of 93 students from 4th to 9th grade, aged 8-14, in public ($n = 61$) and private ($n = 32$) schools were evaluated. Tasks from the AREF instrument, as well as measures of reading comprehension, inhibitory control, cognitive flexibility, working memory, and intelligence, were administered. Correlations between AREF scores and the other measures were performed to assess external construct validity. Performance differences between school groups on AREF subtests were analyzed using ANOVA, t-test, and Mann-Whitney tests, and the internal consistency of the instrument's tasks was evaluated using Cronbach's alpha coefficient.

Results: The scores of the AREF subtests demonstrated significant positive correlations with reading measures (ranging from 0.339 to 0.367) and executive functions (ranging from 0.209 to 0.396). Significant differences were found in the performance of some AREF tasks when comparing individuals from public and private schools, as well as between 4th and 5th graders compared to students in higher grades. The internal consistency of the tasks was low for Graphophonological-Semantic Flexibility (Cronbach's $\alpha = 0.566$), moderate for Inhibitory Control and Flexibility (Cronbach's $\alpha = 0.768$), and high for Working Memory (Cronbach's $\alpha = 0.881$).

Discussion: The results provide initial evidence of construct validity and reliability for the AREF subtests. It is expected that this new neuropsychological test will contribute to the assessment of reading skills and executive functions, assisting in guiding clinical and educational interventions for individuals with and without neurodevelopmental disorders.

KEYWORDS

reading, reading comprehension, executive functions, psychometric validation, inhibitory control, cognitive flexibility, working memory, neuropsychological assessment

1 Introduction

The development of reading and executive functions represents a central area of interest in cognitive research, given its intricate complexity and broad implications for cognitive development (Peng and Kievit, 2020; Burgess and Cutting, 2023). Competence in reading not only stands as a crucial element for academic and professional success but is also imperative for full integration into society (Rabiner et al., 2016; OECD, 2023). However, the acquisition of reading skills is a multifaceted and challenging process, extending beyond mere word decoding to demand equally meaningful comprehension of textual content (Dehaene, 2009; Fonseca et al., 2020).

From a theoretical standpoint, the dual-route cognitive model has often been employed to describe the decoding process in reading, emphasizing the interaction between orthographic-visual analysis and the lexical and phonological routes (Coltheart et al., 2001). However, while this process is fundamental, it proves insufficient to achieve a substantial level of reading proficiency (Kendeou et al., 2014). Therefore, other cognitive processes are also implicated in reading, allowing us to transcend the scope of mere lexical decoding.

One proposal seeking to explain reading comprehension by incorporating decoding into its model is the Simple View of Reading (Gough and Tunmer, 1986). According to this hypothesis, reading comprehension results from Decoding X Linguistic Comprehension, illustrating that reading requires the contribution of both variables for its effectiveness. It is widely accepted that the ability to decode text constitutes a fundamental requirement for comprehension (Perfetti and Hogaboam, 1975). Nevertheless, the Simple View of Reading appears to solely focus on bottom-up processes involved in the activity, rather than presenting a suggestion that includes metacognitive abilities for the reader to assimilate the content of the text (Spencer et al., 2020).

One of the cognitive domains most closely related to effective processing of reading and textual comprehension is executive functions (EFs) (Gonçalves et al., 2017; Follmer, 2018). Executive functions comprise a set of high-level cognitive processes that enable flexible adaptation to diverse contexts, suppression of inappropriate impulsive responses, and temporary maintenance of crucial information in a variety of situations (Diamond, 2013). They are responsible for the regulation and supervision of complex tasks involving planning, decision-making, and problem-solving (Diamond, 2013).

Although there is no consensus regarding the components of executive functions, Miyake et al. (2000) relied on psychometric data to assess the validity of the three-factor model. Following the administration of executive function tests in a sample, confirmatory factor analysis was conducted, which supported the three components: shifting, updating (monitoring and maintaining information in working memory), and inhibition (inhibition of dominant or prepotent responses). The results indicated that, although moderately correlated, the factors are distinct constructs. Diamond (2013) maintains that the three-factor model has been supported in numerous neuropsychological studies, wherein working memory, inhibitory control, and cognitive flexibility comprise the core functions. Working memory refers to the ability to temporarily retain and manipulate information. Inhibitory control consists of the ability to restrain automatic or ongoing behaviors and suppress irrelevant stimuli. Cognitive flexibility, on the other hand, enables adaptation to changes in rules or environmental stimuli, resulting in behavioral adjustments.

Executive functions begin their development in childhood and continue to develop during adolescence, reaching maturity in adulthood (Romine and Reynolds, 2005). Despite this continuous growth, development is not linear, as skills may show more pronounced improvements depending on the period of life and the construct being analyzed (Huizinga et al., 2006). For example, from early childhood, rudimentary behaviors of inhibition, information manipulation, and flexibility are already observable (Capilla et al., 2004). Childhood, in particular, is a crucial period for the rapid development of executive functions, with significant improvement between the ages of 5 and 7, followed by a moderate effect between 8 and 15 years, and a lesser effect between 15 and 17 years (Best et al., 2011).

With the onset of schooling, the development of executive functions occurs simultaneously with the enhancement of reading ability. In the early school years, students learn the basic principles of word decoding and, in subsequent years, automate this skill to eventually comprehend the texts they read (Verhoeven and Perfetti, 2011). However, it is unclear whether reading and executive functions develop independently, without one influencing the trajectory of the other, or bidirectionally, where one ability affects the other through mutually beneficial interactions (Peng and Kievit, 2020). For example, a meta-analysis investigating the relationship between working memory and reading in individuals aged 4 to 80 years demonstrated that this relationship increases with age, suggesting a bidirectional effect between these skills (Peng et al., 2018). However, in Follmer's (2018) meta-analysis examining the relationship between executive functions and reading comprehension from ages 6 to adulthood, the relationship was more pronounced in children than in adults. Regardless of the type of relationship between the developmental trajectories of the two constructs, current evidence supports the hypothesis that executive functions are fundamental for competent reading comprehension processing, directly influencing the ability to extract accurate information from text, interpret meanings, and maintain attentional focus (Butterfuss and Kendeou, 2018).

For example, in the decoding of isolated words, executive functions play an important role in the simultaneous assimilation of their phonological, orthographic, and semantic information (Cartwright, 2007; Varghese and Shanbal, 2024). Similarly, to achieve text comprehension, cognitive flexibility, inhibitory control, and

working memory operate in particular ways. Cognitive flexibility, for example, is related to the ability to modify strategies applied to text reading, as it involves a process that requires planning (Latzman et al., 2010). Inhibitory control, in turn, plays a crucial role in suppressing previously acquired ineffective reading habits (Kieffer et al., 2013) and in inhibiting irrelevant information for text comprehension (Butterfuss and Kendeou, 2018). Finally, working memory plays a recognized role in text comprehension as it supports the retention, manipulation, and association of ideas read (Follmer, 2018). The study by Spencer et al. (2020) emphasized that proficient reading comprehension, as well as the ability to make inferences, requires the reader to manipulate information from multiple sources, including their prior knowledge. These processes demand the use of working memory.

Currently, there are several useful paradigms for assessing reading comprehension, such as those based on response formats like Cloze, Multiple Choice, Open Ended, Retell, and Picture Selection (Collins and Lindström, 2021). Similarly, there is a variety of instruments focused on measuring executive functions, such as the Card Sorting Paradigm (e.g., Wisconsin Card Sorting Test), Continuous Performance Test, Go/No-Go, Hayling and Brixton, Span, and Stroop, among others (Nyongesa et al., 2019). However, there is no instrument that utilizes the interaction between these two constructs to develop a paradigm allowing their simultaneous evaluation, for example, using words and texts to identify both reading skills and executive functions. This goal could be achieved through the application of reading tasks where executive demand progressively increases, so that accurate performance depends on both the recruitment of executive functions and reading ability. The absence of such a tool implies a missed opportunity to assess both constructs in a single battery of tasks, which could lead to greater practicality and efficiency in clinical and educational contexts, as well as differentiated analyses compared to existing paradigms.

In this regard, the development and validation of an assessment battery for the components of executive functions and reading comprehension emerge as a valuable strategy to identify students with deficits in these processes. The AREF - Assessment of Reading and Executive Functions (ALEFE - Avaliação da Leitura e das Funções Executivas) was developed with the purpose of measuring such constructs in students from the 4th to the 9th year of elementary school.

Therefore, the present study aims to verify the psychometric properties of a test constructed to assess reading and executive functions. Our hypothesis is that the AREF test will demonstrate evidence of convergent validity through correlations with already validated tests of reading comprehension and executive functions. Specifically, each AREF subtest (Graphophonological-Semantic Flexibility, Inhibitory Control, Flexibility, and Working Memory) is expected to show correlation with the executive function scores it aims to measure. Given that it is a reading test, we hypothesize that subtest results will exhibit stronger correlations with Verbal IQ than with Performance IQ measures. Additionally, we also hypothesize that the subtests will show good evidence of reliability. Finally, we believe that there will be significant differences in AREF battery performance among different age groups, with superior performances observed in older groups compared to younger ones.

2 Methods

2.1 Sample

The research involved a sample of 93 participants, all Brazilian nationals, Brazilian Portuguese speakers, enrolled from the 4th to the 9th grade of Elementary School. Both public and private school students took part in the research; however, only students from public schools comprised the sample of 4th and 5th graders. The age range of the participants varied from 8 to 14 years, and all of them were selected from two Brazilian states, Espírito Santo (21.5%) and Minas Gerais (78.5%).

Participants were recruited after the researchers contacted the schools. The institutions that showed interest in participating in the research distributed the Consent Terms to be signed by the students' parents. School representatives were instructed not to hand out the terms to students who met at least 1 of the following exclusion criteria: (1) manifesting complaints or indicators of visual, auditory, neurological, behavioral, and/or cognitive impairment; (2) receiving a diagnosis of developmental, language, and/or learning disorders; (3) not being duly enrolled in elementary school; (4) absence, objection, or non-participation in all assessment sessions; (5) reporting difficulties in reading; (6) being in a grade not corresponding to chronological age; and (7) not having the consent form signed by the legal guardian.

2.2 Procedures

The Ethics Committee in Research of the Universidade Federal de Minas Gerais approved the present study. Upon ethical approval, contact was established with elementary schools, both public and private, to obtain the necessary institutional authorization to conduct the research. In accordance with the guidelines established in Resolution No. 196/96 and Resolution No. 466/2012 of the National Health Council of the Ministry of Health, all invited institutions were required to sign an Institutional Assent Form. Once this authorization was obtained, the school staff were informed in advance about the study objectives and the procedures for selecting participants, aiming to gain their support in students' adherence to the research. For the subjects' participation in the study, parents or legal guardians were requested to sign the Informed Consent Form.

The administration of the AREF battery, along with the complementary tests used for validation, was conducted by a team consisting of three psychologists and eight psychology students, all experienced in administering psychometric tests. A criterion was that the researchers responsible for administering the AREF battery were not the same ones who conducted the complementary tests with the same student, ensuring the independence of the assessments.

The administration sessions were scheduled in advance with the schools to ensure that the battery was administered individually to each student. Before starting the test administration, researchers made sure to create a comfortable and age-appropriate environment for the child, with a table and appropriate testing materials.

The administration of the AREF Battery (composed of the Graphophonological-Semantic Flexibility, Inhibitory Control, Flexibility, and Working Memory subtests) and other tests was divided

into two sessions, aiming not to remove the student from the classroom for a single prolonged period. These sessions were spread over two consecutive days, with the entire AREF Battery administered on 1 day and the other tests on the other day (not always in that order). The average duration of each administration session ranged from 30 to 40 min, depending on individual performance and specific needs of each student. It is relevant to highlight that the majority of participants showed interest and engagement in the proposed activities, not expressing fatigue during the assessment process.

After the data collection was completed, individual reports were prepared for each student, detailing their performance on the tasks already commercially available (FDT, WASI, WISC-IV Digits, and PROLEC/PROLEC-SE-R), as a counterpart to the participating institutions. These reports were delivered to the schools with the objective not only to provide access to information about the students' performance but also to understand their individual needs, enabling the planning of personalized educational interventions if necessary.

2.3 Assessment of reading and executive functions

The AREF test consists of 4 subtests, each of them assesses specific aspects of reading comprehension and executive functions: the Graphophonological-Semantic Flexibility task, the Inhibitory Control task, the Flexibility task and the Working Memory task. All of them will be described in the next sessions.

2.3.1 Graphophonological-semantic flexibility

The graphophonological-semantic flexibility plays a crucial role in the ability to comprehend words as it allows for the flexibility of semantic and phonological aspects in word reading. This ability contributes to fluent word reading in early readers (Cartwright et al., 2019). The present study investigated the capacity to switch between graphophonological and semantic components of printed words through an adapted task from previous works (Cartwright, 2007; Cartwright et al., 2010). The resources used in this activity consisted of four sets of cards, including a training set and three test sets. Each set contained 12 cards, each with a printed word, allowing classification along two simultaneous dimensions in a 2×2 matrix, considering both the initial phoneme and the word's meaning. In the exemple set, 12 cards were presented to the student with the instruction:

"I have here some cards for you to organize. You can sort them in two ways simultaneously: by their initial letter and by their meaning." The administrator would take the first card, show the word to the participant, and continue:

"*See, I will place the word MOOSE, which is an animal, up here.*" The word was placed in the upper left quadrant. A new card was taken out and its word was shown. "*The word CAMEL is an animal too; so, I will also place it at the top, like MOOSE, but I cannot place it on the left side, because this side is for words with M, so I will place it here on the right. Note that I cannot place words of the same meaning, representing the same category like ANIMALS, diagonally.*" The administrator would take the next card, show the word to the participant, and continue the instruction: *"The next word is CHERRY. Since it starts with C, I will place it on the right side, like CAMEL, but I cannot place it on the top because I only put animals there, so I will place it here at the bottom. Note that I also cannot place words with the same letter diagonally."* A new word was shown to the participant [MANGO], and the administrator would ask: "*Where do I place this next word?*" It was expected that the participant would indicate, either physically or verbally, the bottom left corner, corresponding to the row where fruits are and the column where words with M are. If they gave the correct answer, the administrator would congratulate them and ask why they chose that space.

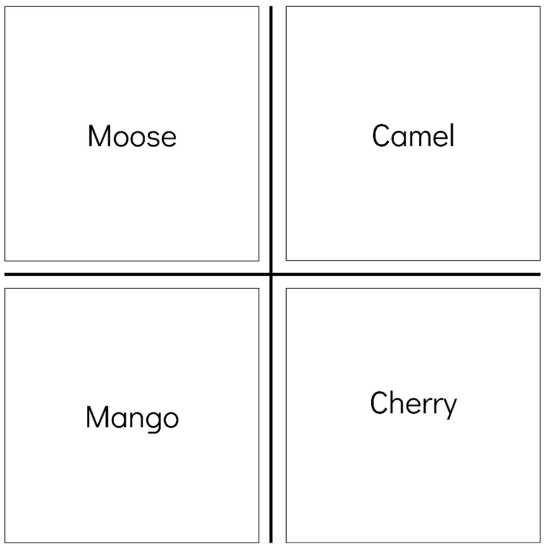


FIGURE 1
Example of classifying 4 words in the Graphophonological-Semantic Flexibility subtest matrix, simultaneously considering the meaning of the words and the initial letters.

In the justification, it was expected that the student's response would encompass the division of words in the matrix, simultaneously considering the initial letter and the meaning. For example: "At the top I put the animals and at the bottom the fruits. On the left side, I placed the words with M and on the right side the words with C". Figure 1 shows an example of a possible classification expected in this task, in which on the left side of the matrix there are only words starting with the letter M, on the right side there are only words starting with the letter C, on the top there are only words animals, and in the lower part only the fruits.

After the administrator classified the first 3 words, explaining the rule, and after the participant classified the fourth word, the administrator would hand over the other 8 cards from this set for the student to perform the classifications. When the student performed the task correctly, they were congratulated, and when the execution was incorrect, the administrator would say "*Not quite*" and reinforce the classification rule.

After the training set, the first test set was conducted, preceded by the instruction: "*Very well. Now I will give you other words, and you will separate them the same way we did until now: by the letter and the meaning. The letters and meanings will not always be the same as the ones we did until now, but you can separate them the same way. If you make a mistake and want to change a word, continue the activity, and you can change the word's place at the end. You may begin.*" After classifying this set, the administrator would ask why the participant organized the words that way. Again, it was expected that they would

respond that they considered, simultaneously, the initial letters and the meaning of the words.

Results were recorded in terms of the time required to classify each set of words, along with the assembly of the matrix (1 point when correct and 0 when incorrect), followed by justifications for their classifications (2 points when correct and 0 when incorrect). Once the test items were scored, the administrator did not provide feedback on the participant's performance.

To ensure the standardization of the test application, each set of cards was always presented in the same sequence, specifically test sets 1, 2, and 3. Additionally, the sequence of words within each set remained constant throughout the study.

Scoring followed the following criteria: one point was awarded for the accurate assembly of the matrix, and two points were assigned for an adequate justification of the process performed. Considering that there were 3 items scored, the maximum score obtained in this subtest was 9. In cases where there was an error in both stages of the activity, the score was null.

The study took into account the characteristics of the language used in the stimuli, such as high-frequency orthographic words in Brazilian Portuguese. This selection was conducted using data repositories available on the platforms http://lexicodoportugues.com/ and https://www.corpusdoportugues.org/now/. The selection criteria included syllabic length (disyllabic, trisyllabic, and polysyllabic) and the complexity of words according to the structure of the initial syllable. Polysemous, homographic, and monosyllabic words were deliberately excluded. The mentioned guidelines ensured the diversity of the chosen words by varying in regularity, length, and syllabic complexity.

2.3.2 Inhibitory control

The inhibitory control subtest assesses the student's ability to suppress automatic responses and resist distractions during reading. Divided into three stages (1 baseline and 2 inhibition), the participant is instructed to read narrative texts aloud, all consisting of 94 words each. After each reading, the participant must retell the events in the story and orally answer three specific questions about the text. The answers are definitive and require the direct retrieval of the information read, without the need for inferences from secondary information. The retelling involves identifying eight specific events, which are recorded in the response booklet as a checklist. Each event remembered in the retelling is counted as 1 point, as well as each correct answer. The maximum possible score for each stage of the subtest is 11 points, and the reading time for each text is timed.

The first stage consists of a typical text without interference from other colors, as illustrated in Figure 2, designed to assess reading fluency and text comprehension, measured, respectively, by the reading time and the score obtained from the retelling and responses to the questions. The results of this stage are used as a baseline for comparison with the results of inhibitory control and flexibility. Therefore, this task is called Baseline Text (BT).

The second stage begins with a preliminary training task, where the participant is presented with a text containing lines in three different colors: blue, red, and black, as shown in Figure 3. The participant is instructed to read aloud only the black sentences. After reading this training text, the evaluation text is displayed, and the instructions are reiterated. The reading time is timed in seconds from the start of reading and is stopped upon completion. The time is recorded in the response booklet, as well as the number of incorrectly read words, the number of colored sentences read, and the number of black lines ignored. After this stage, the participant is again invited to retell the story and orally answer three questions about the text read, and the score is recorded in the booklet based on the participant's performance. This is the first task that seeks to evaluate Inhibitory Control (IC-1).

In the third stage, a text with lines of different colors is presented once again, serving as further training before the official task. The examiner instructs the participant to continue reading only the sentences in black, avoiding reading sentences in other colors. Additionally, there are black words within colored sentences, which should not be read, as shown in Figure 4. The participant must only read aloud the sentences in which all words are black. Similar to the previous stages, the reading time is timed in seconds from the beginning and stopped upon completion. The time is recorded in the response booklet. Accuracy is also recorded, noting any instances where the participant read words that should have been ignored, including errors of reading black words within colored sentences and any failures involving reading lines in colors other than black. After reading, the participant is asked to retell the story and orally answer three questions about the text read. Both correct and incorrect responses are recorded in the response booklet. This is the second task that seeks to evaluate Inhibitory Control (IC-2).

2.3.3 Flexibility

In the Flexibility (FL) subtest, the participant is required to alternate reading sentences of different colors according to a

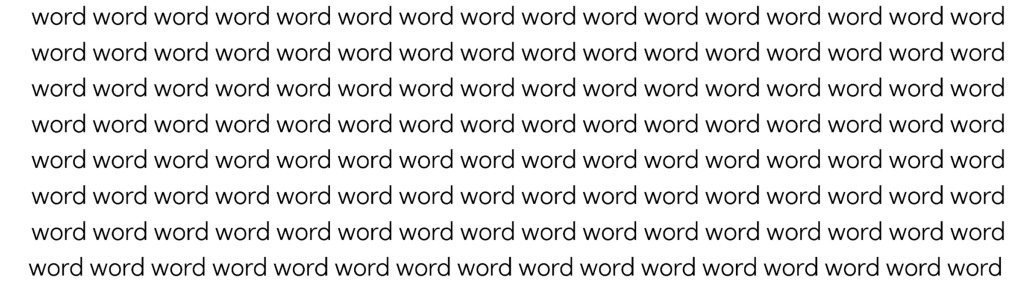


FIGURE 2
Example of text without interference from other colors used as a baseline for the AREF Inhibitory Control and Flexibility subtests.

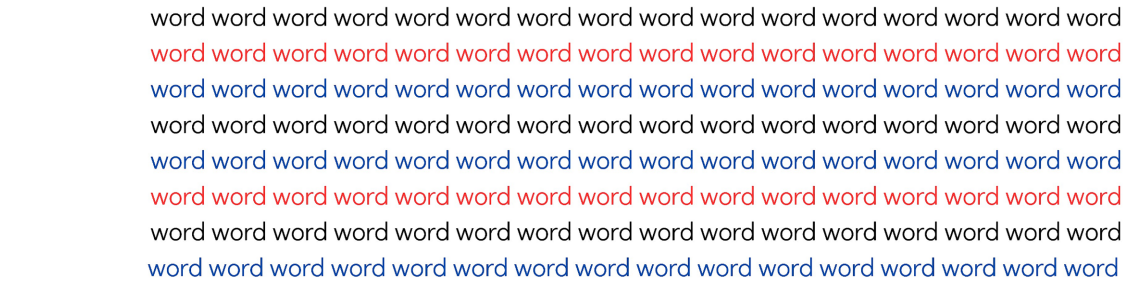


FIGURE 3
Example of text from the first Inhibitory Control task of the AREF test, in which the participant must read only the black lines.

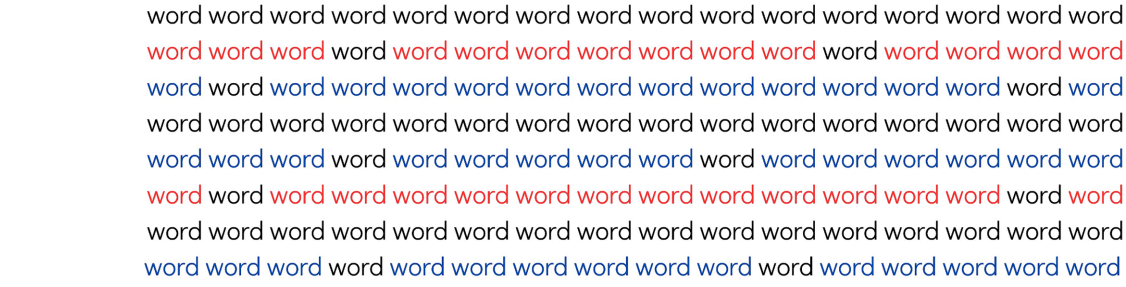


FIGURE 4
Example of text from the second Inhibitory Control task of the AREF test, in which the participant must read only the black lines while avoiding reading words written in black on colored lines.

FIGURE 5
Example of text from the AREF Flexibility task, in which the participant must alternate reading between sentences of different colors depending on the color of the visual sign (line) present in the text.

visual cue, aiming to assess the schoolchild's cognitive flexibility. At the beginning of this task, before the sentences that comprise the text, there is a continuous black line, as shown in Figure 5. This black line indicates that the participant should read only the black sentences, ignoring the red or blue sentences. The reading of the black sentences should continue until the appearance of another visual cue indicating a change in the color of the sentences to be read. In Figure 5, as in the original task, this cue is represented by a red line, after which the participant should read only the red sentences. The reading of the red sentences should continue until a new visual cue indicates a change in color. In Figure 5, as in the original task, this cue is the second black line. The test is preceded by a training item. The reading time for this subtest is recorded, followed by the retelling of the story and responses to three specific questions. Responses are scored based on the direct retrieval of the information read, with a total of 8 events to be identified during the retelling, similar to the baseline text and the inhibitory control tasks.

2.3.4 Working memory

In the Working Memory subtest, the examiner instructs the participant to read aloud sentences and, after reading, to retell the story in reverse order, without the visual aid of the text. Initially, a practice session with a text consisting of 2 events is conducted, followed by the commencement of the evaluative task. An example of a stimulus text for practice and the expected response after reading is as follows: The stimulus text 'I went to the park. I played soccer.' is provided, and the expected response is 'I played soccer. Prior to that, I went to the park'.

In total, after the practice session, participants were provided with seven different texts to read aloud. These narrative texts, which contained words commonly used in Brazilian Portuguese, varied in content and in the number of events included. The first text presented three events, and each subsequent text introduced one additional event compared to its predecessor, thereby gradually increasing the demand for information retrieval (span). Each sentence in the texts, ending with a period, represents an event to be remembered. Each event remembered correctly corresponds to one point. However, once an event is remembered and reported, any other event will only be scored if it chronologically preceded it.

In composing the original texts in the Working Memory subtest, the following criteria were applied: segmentation of events through periods and ensuring consistency in the length of sentences.

2.4 Neuropsychological protocol

To establish the construct validity of the instrument developed to concurrently assess reading and executive functions, analyses were conducted to verify evidence of external construct validity. For this purpose, the following instruments were used: the Vocabulary and Matrix Reasoning subtests of the WASI (Wechsler, 2014); the PROLEC Text Comprehension (Capellini et al., 2012) for 4th and 5th-grade students; the PROLEC-SE-R Narrative Comprehension for 6th to 9th-grade students (Cuetos et al., 2022); the Five Digit Test (FDT) (Sedó et al., 2015); and the Digit Span subtest of the WISC-IV (Wechsler, 2013).

2.4.1 Wechsler abbreviated scale of intelligence

To assess general intelligence across a wide age range, the Wechsler Abbreviated Scale of Intelligence (WASI) was utilized, comprising the Vocabulary and Matrix Reasoning subtests. Whereas the Vocabulary subtest assesses verbal comprehension and knowledge of word meanings, Matrix Reasoning evaluates nonverbal fluid reasoning through visual patterns.

Individuals who scored below 70 on the IQ test were excluded, leading to the elimination of one participant.

2.4.2 PROLEC's text comprehension

The PROLEC assesses reading processes in children from 2nd to 5th grade of elementary school. The subtest consists of four brief texts, followed by questions addressing both literal and inferential aspects of textual comprehension. Each text has 4 questions, totaling 16 questions distributed among the texts. A score of 1 point is assigned to each correct answer, while incorrect answers receive 0 points, allowing participants to obtain a maximum of 16 points.

2.4.3 PROLEC-SE-R's narrative comprehension

To assess narrative reading comprehension in later grades, the PROLEC-SE-R was employed, targeting students from 6th to 9th grade of elementary school and from 1st to 3rd grade of high school. This instrument involves reading a narrative text followed by 10 multiple-choice questions, with the allowance to consult the text during the assessment.

2.4.4 Five digit test

Another instrument employed was the Five Digits Test. This test assesses inhibitory control and cognitive flexibility. It comprises four distinct stages. In the first stage, named Reading, participants are presented with rectangles containing numerals from 1 to 5, with the quantity of numerals inside the rectangle corresponding to the magnitude of the represented number (e.g., two numerals inside the rectangle for the number 2). The objective is for the participant to name the numerals contained in 50 stimuli as quickly as possible. In the second stage, Counting, the rectangles contain up to five asterisks, and participants must count the quantity of asterisks in 50 stimuli as quickly as possible. The third stage, called Choosing, repeats the presentation of the rectangles, but this time with an incongruent condition, meaning the quantity of numerals inside the rectangle does not match the magnitude of the number (e.g., three numerals 4 inside the rectangle). Participants must count the quantity of numerals in 50 stimuli as quickly as possible, inhibiting the automatic response of pronouncing the name of the represented numeral. In the fourth stage, Shifting, participants continue counting the quantity of numerals, but when presented with a rectangle with a thicker border, they must say the name of the numeral contained. Thus, counting and naming responses are alternated. Also, 50 stimuli are presented in this stage. In addition to the four mentioned stages, the test provides measures of inhibition and flexibility, derived from the time spent in Stages 1, 3, and 4. The inhibition measure is calculated by subtracting the time from Stage 1 (Reading) from the time from Stage 3 (Choosing). The flexibility measure is calculated by subtracting the time from Stage 1 (Reading) from the time from Stage 4 (Shifting). Test correction considers the total time taken for each stage, as well as the quantity of errors made.

2.4.5 Digit span subtest (WISC-IV)

The "Digit Span" subtest of the WISC-IV was employed to assess working memory and auditory attention in children. In this subtest, the examiner presents a series of digits for the participant to repeat either in the same order (Forward) or in reverse order (Backward), with a gradual increase in difficulty.

2.5 Data analysis

All analyses were carried out considering the total results of each subtest of AREF. For the GSF, IC, and FL tasks, in which the duration of execution was measured, the final score of each subtest was calculated considering both accuracy (number of correct responses) and time taken. This approach is supported by evidence in the literature indicating that, in both executive function and reading tests, time is a crucial variable for predicting performance. For example, Magnus et al. (2019) demonstrated that the joint use of accuracy and reaction time improves the precision of inhibitory control

measurement compared to models that use only accuracy. Similarly, Su and Davison (2019) noted that including response time measures can improve the validity of a reading test, as lower response times during reading are observed in individuals with higher ability. Therefore, we opted to include both accuracy and time in our scoring approach to ensure a more accurate and valid assessment of performance.

Performance on the GSF task was evaluated by summing the Execution Points (EP) and Justification Points (JP) of the three items composing the task, multiplied by 60, and divided by the sum of the time (T) of the three items. This evaluation resulted in the efficiency score (GSF-ES) in task execution, as demonstrated by the formula below:

$$GSF\ Efficiency\ Score = \left(EP1 + JP1 + EP2 + JP2 + EP3 + JP3\right) x\ 60 / \left(T1 + T2 + T3\right)$$

Regarding AREF's Inhibitory Control and Flexibility subtests, statistical analyses were also conducted considering the efficiency score obtained in each activity, using the following calculation: (Recall points + response points) x 60 / reading time in seconds. As a result, there were four efficiency scores in that stage: from Baseline Text (BT-ES), from the first Inhibitory Control task (IC1-ES), from the second Inhibitory Control task (IC2-ES), and from Flexibility task (FL-ES).

The result of the Working Memory task (WM Total) was calculated by summing the results of the seven items composing the task:

$$WM\ Total = WM1 + WM2 + WM3 + WM4 + WM5 + WM6 + WM7$$

To verify evidence of convergent validity, it was examined the relationship between AREF subtests results and external measures with correlation analysis. Prior to this analysis, the multivariate Shapiro–Wilk test was applied, indicating that the joint distributions of the variables were non-parametric, justifying the use of Spearman correlation.

For correlation analysis, participants' results on external measures were transformed into scores or ratings obtained from the respective tasks. Regarding WASI, T-scores of the applied subtests (Vocabulary and Matrix Reasoning) were utilized. Regarding the FDT, inhibition and flexibility percentiles were used. As for the WISC-IV Digit Span subtest, both forward and backward span, as well as Scaled Scores (SS) obtained throughout the task, were employed. Concerning the reading comprehension subtests of the PROLEC and PROLEC-SE-R tests, their classifications based on individual performance had to be unified. In PROLEC, administered in the 4th and 5th grades, the categories are "SD" (Severe Difficulty), "D" (Mild Difficulty), and "A" (Average), whereas in PROLEC-SE-R, administered from the 6th to 9th grades, the categories include "SD" (Severe Difficulty), "D" (Mild Difficulty), "L" (Low), "A" (Average), and "H" (High). As our analyses involved the entire population from the 4th to 9th grades, the "Low," "Average," and "High" categories from PROLEC-SE-R were grouped into a single category, corresponding to the "Average" classification of PROLEC. This approach was adopted to standardize categories and ensure greater precision in statistical analyses involving both population groups.

Given that AREF subtests measure distinct constructs, some observations are needed. Firstly, the correlations conducted for the GSF efficiency score were the same for Inhibitory Control and Flexibility tasks, which included reading comprehension measures (PROLEC and PROLEC-SE-R subtests), executive function measures (Flexibility and Inhibition percentiles of the FDT), and verbal and performance measures of the WASI (Vocabulary subtest and Matrix Reasoning subtest, respectively). Secondly, correlations of WM Total were performed with the same aforementioned reading comprehension measures, Working Memory measures (forward and backward span, in addition to the Scaled Scores), and verbal and performance measures of the WASI as well.

In this study, it was also investigated the differences between the mean performance on the Baseline Text task and the other tasks of the Inhibitory Control and Flexibility subtests. Before comparison, the Shapiro–Wilk test was applied to check the data distribution. In cases where the distribution was non-parametric, the Wilcoxon test was used for comparison, while the effect size was evaluated by the Point-Biserial Correlation Coefficient. When the distribution was parametric, the paired t-test was employed, with the effect size calculated by Cohen's d test. The comparisons made were between the group's efficiency performance in the Baseline Text and efficiency in IC-1, IC-2, and FL. These comparisons were feasible because all tests shared the same efficiency calculation and the same format, including the same number of words in the target texts and the same amount of clauses to be retold and questions to be answered.

To strengthen the evidence of construct validity, an investigation was conducted on potential differences in the performances of distinct groups. The instrument's ability to differentiate these groups provides evidence of concurrent validity, which is used to evaluate test-criterion relationships (American Educational Research Association, 2014), where the scores obtained on the tasks predict outcomes observed at the time of test administration.

Considering that executive function and reading comprehension skills improve throughout schooling, differences in the performance of individuals from different school years on the AREF subtests were measured. To evaluate the performance of groups from different school years on the AREF subtests, Analysis of Variance (ANOVA) was applied. Detailed group comparison analysis was conducted only when the ANOVA indicated statistical significance ($p < 0.05$), meaning significant differences were detected between the groups. In such cases, the Levene's test was employed to examine the homogeneity of variances among the groups. If Levene's test revealed a $p > 0.05$, a *Post Hoc* analysis using Tukey's test was performed to identify which groups showed significant differences in performance.

Given that previous studies have identified significantly different performances between students from public and private schools in reading comprehension (Marques de Oliveira et al., 2024; Cáceres-Serrano and Alvarado-Izquierdo, 2017; Çigdemir and Akyol, 2022) and executive functions (Jacobsen et al., 2017), the AREF scores of participants from both school types were compared. To perform this comparison, the normality of the data distribution was first assessed using the Shapiro–Wilk test. If the distribution was non-parametric, the Mann–Whitney test was employed. When parametric distribution was confirmed, Levene's test was used to evaluate the equality of variances. In cases where Levene's test did not show significance, the independent samples t-test was subsequently applied, with effect size estimated using Cohen's *d*. As previously mentioned, the sample of

10.3389/fpsyg.2024.1399388

4th and 5th-grade students consisted exclusively of public school students. To eliminate the possibility that differences in performance between public and private schools were due to the younger average age of public school students, the comparison between school types was conducted only for students from 6th to 9th grade (N = 50). The identification of performance differences between students from public and private schools on the AREF test also contributes as evidence of concurrent validity.

The internal consistency of each AREF subtest was assessed using Cronbach's alpha coefficient. It should be noted that, once Inhibitory Control subtest and Flexibility subtest were made up of a similar structure (recall points, response points and reading time), and, besides that, required almost the same cognitive constructs, these subtests were grouped in this internal consistency analysis.

All statistical analyses were performed using JASP 0.17.2.0 software (JASP Team, 2023).

3 Results

The characteristics of the participants are presented in Table 1, which includes information on age, gender, grade level, and the type of school within the collected sample.

3.1 Construct validity

Table 2 illustrates the Spearman correlation of efficiency scores obtained in the Graphophonological-Semantic Flexibility subtest with the classification of results from the PROLEC and PROLEC-SE-R subtests, along with the percentiles of inhibition and flexibility from the FDT, and the T-scores of the Vocabulary and Matrix Reasoning subtests of the WASI.

The results indicate that the efficiency score obtained in the GSF task presented weak, but significant, positive correlations with the FDT Inhibition percentile [rs (93) = 0.209; $p = 0.045$]. Likewise, the correlations of the GSF subtest efficiency scores were positive and significant with the classification obtained in the PROLEC and PROLEC-SE-R tests, of moderate magnitude [rs (91) = 0.355; $p < 0.001$], as well as with the T-score of the WASI Vocabulary subtest [rs (91) = 0.348; $p < 0.001$]. No significant correlations were found between the GSF subtest and the FDT Flexibility percentile [rs (93) = 0.116; $p = 0.266$] and the WASI Matrix Reasoning T-score [rs (93) = 0.109; $p = 0.298$].

Table 3 depicts the Spearman correlation of scores from the Inhibitory Control and Flexibility subtests of the AREF with the classifications of results from the PROLEC and PROLEC-SE-R subtests, along with the percentiles of inhibition and flexibility from the FDT, and the T-scores of the Vocabulary and Matrix Reasoning subtests of the WASI.

As expected, the efficiency scores in all AREF texts showed a positive and significant correlation, of moderate magnitude, with the classification of performance in the PROLEC and PROLEC-SE-R subtests (ranging from 0.339 to 0.367). The AREF scores also showed positive and significant correlations, of weak to moderate magnitude, with the FDT Inhibition percentile (ranging from 0.284 to 0.387), as well as with the WASI Vocabulary subtest T-score (ranging from 0.262 to 0.412). Only the first inhibitory control task of the AREF test showed a positive and significant correlation, of weak magnitude, with the FDT flexibility percentile [rs (93) = 0.265; $p < 0.010$]. There was no significant correlation between the AREF subtests and the T-score of the Matrix Reasoning subtest of the WASI.

The results showed that the total score of the AREF working memory task correlated significantly and positively with the classification of the PROLEC and PROLEC-SE-R subtests [moderate magnitude, rs (93) = 0.365; $p < 0.001$], with the Scaled Scores of the WISC-IV Digit Span subtest (weak magnitude, rs (93) = 0.259; $p = 0.012$), with the direct span of the WISC-IV Digit Span subtest

TABLE 1 Characterization of participant profiles ($n = 93$) according to age range, gender, grade level and type of school.

Feature	Category	No.	%
Age	8 years	1	1,1%
	9 years	17	18,3%
	10 years	26	28%
	11 years	12	12,9%
	12 years	14	15,1%
	13 years	14	15,1%
	14 years	9	9,7%
Gender	Male	37	39,8%
	Female	56	60,2%
Grade level	4th grade	17	18,3%
	5th grade	26	28,0%
	6th grade	15	16,1%
	7th grade	16	17,2%
	8th grade	13	14,0%
	9th grade	6	6,5%
Type of school	Public	61	65,6%
	Private	32	34,4%

TABLE 2 Spearman correlation of the efficiency of the graphophonological-semantic flexibility subtest of AREF with the classification of PROLEC and PROLEC-SE-R and with the percentiles of inhibition and flexibility of FDT.

Variable		GSF-ES
PROLEC classification	Spearman	0.355
	p-value	<0.001***
Inhibition - PC	Spearman	0.209
	p-value	0.045*
Flexibility - PC	Spearman	0.116
	p-value	0.266
WASI vocabulary	Spearman	0.348
	p-value	<0.001***
WASI matrix reasoning	Spearman	0.109
	p-value	0.298

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. PROLEC Classification, Classification of PROLEC and PROLEC-SE-R. Inhibition – PC, Percentile data of inhibition from FDT. Flexibility – PC, Percentile data of flexibility from FDT. GSF-ES, Efficiency score of the graphophonological-semantic flexibility subtest.

TABLE 3 Spearman correlation of the efficiency of tasks from the baseline text, the inhibitory control and the flexibility subtests of AREF with the T score of the vocabulary and matrix reasoning subtests of WASI and the percentile of inhibition and flexibility from FDT.

Variable		BT-ES	IC1-ES	IC2-ES	FL-ES
PROLEC classification	Spearman	0.339	0.367	0.339	0.358
	p-value	<0.001***	<0.001***	<0.001***	<0.001***
Inhibition - PC	Spearman	0.300	0.387	0.284	0.358
	p-value	0.004**	<0.001***	0.006**	<0.001***
Flexibility - PC	Spearman	0.112	0.265	0.203	0.100
	p-value	0.286	0.010*	0.051	0.341
WASI vocabulary	Spearman	0.307	0.412	0.390	0.262
	p-value	0.003**	<0.001***	<0.001***	0.011*
WASI matrix reasoning	Spearman	0.131	0.172	0.124	0.185
	p-value	0.211	0.099	0.236	0.076

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. PROLEC Classification, Classification of PROLEC and PROLEC-SE-R. Inhibition – PC, Percentile data of inhibition from FDT. Flexibility – PC, Percentile data of flexibility from FDT. TB-ES, Baseline text efficiency. IC1-ES, Efficiency of Text 1 from inhibitory control subtest. IC2-ES, Efficiency of Text 2 from inhibitory control subtest. FL-ES, Efficiency of the flexibility subtest.

TABLE 4 Spearman correlation of the score obtained in the working memory subtest of AREF with the result classifications of PROLEC and PROLEC-SE-R subtests, with the digit span scaled scores, of the WISC-IV, and with the T-score of the vocabulary and matrix reasoning subtests of the WASI.

Variable		WM total
PROLEC classification	Spearman	0.365
	p-value	<0.001***
Digit span - SS	Spearman	0.259
	p-value	0.012*
Forward span	Spearman	0.396
	p-value	<0.001***
Backward span	Spearman	0.160
	p-value	0.125
WASI vocabulary	Spearman	0.328
	p-value	0.001**
WASI matrix reasoning	Spearman	0.241
	p-value	0.020*

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. WM Total, total points obtained in the 7 items of the working memory task of AREF. Digits – SS, Scaled scores data of digit span from WISC-IV.

(moderate magnitude, rs (93) = 0.396; $p < 0.001$), and with the T-score of the Vocabulary subtests (moderate magnitude, rs (93) = 0.328; $p < 0.001$) and Matrix Reasoning [weak magnitude, rs (93) = 0.241; $p = 0.020$] from WASI. These data are presented in Table 4.

The comparison of efficiency between the Baseline Text and IC-1 Text was conducted using the Paired Wilcoxon Test, due to the non-parametric distribution. For other comparisons with parametric distribution, independent samples t-tests were employed. No significant difference was observed when comparing the performance in the Baseline Text to the IC-1 Text (U = 2015.000; $p = 0.898$) or to the FL Text [t (91) = 1.849; $p = 0.068$]. However, a significant difference was identified [t (91) = 2.098; $p = 0.039$] between the performance in the Baseline Text (M = 10.2, SD = 5.5) and the performance in the IC-2 Text (M = 9.2, SD = 5.8), with a small effect size (d = 0.218).

Regarding the performance analyses of different school years, the ANOVA results revealed a significant group effect on all subtests. Effects were observed in the GSF subtest, F (5, 89) = 8.115, $p < 0.001$, $\eta^2 = 0.318$, as well as in IC-1, F (5, 89) = 10.898, $p < 0.001$, $\eta^2 = 0.385$, and in IC-2, F (5, 89) = 195.484, $p < 0.001$, $\eta^2 = 0.312$. Significant differences were also observed in group performance in the Flexibility subtest, F (5, 89) = 120.331, $p < 0.001$, $\eta^2 = 0.242$, and in the WM subtest, F (5, 89) = 10.345, $p < 0.001$, $\eta^2 = 0.373$. In all tasks, significant group differences occurred in most comparisons between students from the 4th and 5th grades and those from other school years. The results of the Analysis of Variance are presented in Table 5, and the comparisons of the different school years in each of the subtests are shown in Tables 6–10.

Regarding the comparison between public and private schools, it was found that the data distributions in GSF, IC-2 and WM were parametric and exhibited equal variances. Therefore, the comparison between the groups was conducted using the Student's t-test. A significant difference in WM performance was observed between the groups [t (48) = −2.135; $p = 0.038$], with a medium effect size indicated by Cohen's d of −0.629, showing higher performance by students from private schools (Mean = 29.3, Standard Deviation = 5.96) compared to those from public schools (Mean = 25.5, Standard Deviation = 6.51). No significant difference was found in GSF performance [t (48) = −0.792; $p = 0.433$] or IC-2 performance [t (48) = −1.477; $p = 0.146$]. The detailed results are presented in Table 11, with the magnitude of the means described in Table 12. Supplementary Figure S1, illustrates the comparison of the average performance of individuals from the two groups in the WM task.

In contrast to the other subtests, the analysis of efficiency score distributions for IC-1 and FL between public and private schools revealed them to be non-parametric. Therefore, Mann–Whitney tests were applied for the analyses. The results indicated statistically significant differences in both IC-1 task (w = 187.500, $p = 0.043$) and FL task (w = 190.000, $p = 0.049$). Rank-Biserial correlations showed medium effect sizes of −0.349 for IC-1 and -0.340 for FL. Participants from private schools demonstrated higher performance compared to those from public schools in the IC-1 task (Mdn private = 13.605 vs. Mdn public = 11.325), as well as in the FL task (Mdn private = 12.700 vs. Mdn public = 9.575). The analysis results are presented in Table 13, and the medians for each group in each subtest are shown in Table 14. Supplementary Figures S2, S3, illustrate, respectively, the comparison of students' performance between school types in the IC-1 and FL tasks.

3.2 Reliability

The AREF's reliability of each subtest was measured by Cronbach's alpha.

Regarding Graphophonological-Semantic Flexibility task, its internal consistency was low (0.566), as indicated in Table 15.

Item-rest correlation of Graphophonological-Semantic Flexibility subtest is presented in Table 16. The points obtained by appropriate locations in the matrix as well as those obtained by correct justifying

TABLE 5 Results of the analysis of variance (ANOVA) for comparison of different school grades in relation to AREF subtests.

Cases	Sum of scores	df	Mean of scores	*f*	*p*	η^2
Grade - GSF efficiency	65.670	5	13.134	8.115	<0.001***	0.318
Residuals	140.801	87	1.618			
Grade - IC-1 efficiency	777.688	5	155.538	10.898	<0.01**	0.385
Residuals	1241.629	87	14.272			
Grade - IC-2 efficiency	977.421	5	195.484	7.888	<0.001***	0.312
Residuals	2155.993	87	24.782			
Grade - FL efficiency	601.656	5	120.331	5.540	<0.001***	0.242
Residuals	1889.571	87	21.719			
Grade - total WM	2505.078	5	501.016	10.345	<0.001***	0.373
Residuals	4213.653	87	48.433			

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. F represents the F statistic of the ANOVA, and p denotes the significance value associated with the analysis. GSF Efficiency, Efficiency of the graphophonological-semantic flexibility subtest. IC-1 Efficiency, Efficiency of the first text of the inhibitory control subtest. IC-2 Efficiency, Efficiency of the second text of the inhibitory control subtest. FL Efficiency, Efficiency of the flexibility subtest. Total WM, Total points obtained in the 7 items of the AREF working memory task.

TABLE 6 Results of *post hoc* comparisons of analysis of variance (ANOVA) comparing the performance of different school years (4th to 9th grade) on the GSF subtest of AREF.

Grade		Mean difference	SE	*t*	p_{tukey}
4	5	0.084	0.397	0.211	1.000
	6	−1.553	0.451	−3.447	0.011
	7	−1.285	0.443	−2.900	0.052
	8	−1.918	0.469	−4.092	0.001
	9	−1.804	0.604	−2.986	0.041
5	6	−1.637	0.412	−3.970	0.002
	7	−1.369	0.404	−3.387	0.013
	8	−2.002	0.432	−4.633	<0.001
	9	−1.888	0.576	−3.276	0.018
6	7	0.268	0.457	0.587	0.992
	8	−0.365	0.482	−0.756	0.974
	9	−0.250	0.615	−0.407	0.999
7	8	−0.633	0.475	−1.333	0.766
	9	−0.519	0.609	−0.852	0.957
8	9	0.114	0.628	0.182	1.000

P-value adjusted for comparing a family of 6.

had a weak positive correlation with the total score on the other items. The time measures, on the other hand, exhibited negative correlations with the total score ranging from weak to moderate.

Concerning Inhibitory Control and Flexibility subtestes, their internal consistency was acceptable (0.768), as shown by Table 17.

In Table 18 are indicated item-rest correlation of Inhibitory Control and Flexibility subtests. The correlations of punctuations obtained by story retelling and question answering with the total score were positive, ranging from weak to moderate. In relation to reading times, their correlations with total score were negative, ranging from moderate to high.

Regarding the Working Memory subtest, Cronbach's Alpha showed a high internal consistency (0.881), as illustrated in Table 19.

The item-rest correlation of Working Memory subtest is reported in Table 20. It revealed positive correlations with total score, ranging from moderate to high.

4 Discussion

The primary goal of this article was to furnish evidence regarding the construct validity and reliability of a new neuropsychological test designed to evaluate both executive functions and reading comprehension. Convergent validity was indicated by correlation results, concurrent validity was verified by the prediction of outcomes at the time of task performance (school year and type of school) and reliability was measured by internal consistency.

The results evidenced satisfactory psychometric qualities of the constructed tasks, manifested by significant and positive correlations with external measures of executive functions and reading comprehension, as well as adequate internal consistency of the AREF tasks. The GSF subtest showed expected correlations with reading measures, executive functions, and the Verbal IQ T-score of the WASI verbal IQ task. Although these correlations were weak, they are aligned with initial expectations, suggesting that graphophonological-semantic flexibility may serve as a relevant indicator of reading comprehension, corroborating previous findings by Cartwright (2007), Cartwright et al. (2010), and Varghese and Shanbal (2024). Additionally, it was observed that the inhibition measure of the FDT test correlated significantly with the GSF task, while the flexibility measure did not show correlation. This result can be interpreted in light of previous studies indicating that inhibition is a process that precedes flexibility (Diamond, 2013).

In the two Inhibitory Control subtests, significant correlations were identified with the reading comprehension measures (PROLEC and PROLEC-SE-R), inhibition percentile obtained in the FDT and the verbal IQ measure. The convergence between the results of the AREF Inhibitory Control tasks and the external reading measures indicate that the two share the same required construct, namely reading comprehension. The correlations between the results of the AREF Inhibitory Control subtests and the inhibition measure of the FDT align with our initial hypothesis that these relationships would

TABLE 7 Results of *post hoc* comparisons of analysis of variance (ANOVA) comparing the performance of different school grades (4th to 9th) on the IC-1 subtest of AREF.

Grade		Mean difference	SE	*t*	p_{tukey}
4	5	−0.185	1.178	−0.157	1.000
	6	−5.396	1.338	−4.032	0.002
	7	−5.338	1.316	−4.057	0.001
	8	−6.133	1.392	−4.406	<0.001
	9	−7.480	1.794	−4.170	<0.001
5	6	−5.212	1.225	−4.255	<0.001
	7	−5.153	1.200	−4.293	<0.001
	8	−5.948	1.283	−4.635	<0.001
	9	−7.295	1.711	−4.264	<0.001
6	7	0.058	1.358	0.043	1.000
	8	−0.737	1.432	−0.515	0.995
	9	−2.084	1.825	−1.142	0.862
7	8	−0.795	1.411	−0.564	0.993
	9	−2.142	1.808	−1.184	0.843
8	9	−1.347	1.865	−0.722	0.979

P-value adjusted for comparing a family of 6.

TABLE 8 Results of *post hoc* comparisons from analysis of variance (ANOVA) comparing the performance of different school grades (4th to 9th) on the IC-2 subtest of AREF.

Grade		Mean difference	SE	*t*	p_{tukey}
4	5	−1.229	1.553	−0.791	0.968
	6	−5.325	1.763	−3.020	0.038
	7	−7.380	1.734	−4.256	<0.001
	8	−7.964	1.834	−4.342	<0.001
	9	−7.941	2.364	−3.359	0.014
5	6	−4.096	1.614	−2.538	0.125
	7	−6.151	1.582	−3.889	0.003
	8	−6.736	1.691	−3.983	0.002
	9	−6.713	2.255	−2.977	0.042
6	7	−2.055	1.789	−1.149	0.859
	8	−2.640	1.886	−1.399	0.727
	9	−2.616	2.405	−1.088	0.885
7	8	−0.585	1.859	−0.315	1.000
	9	−0.561	2.383	−0.236	1.000
8	9	0.023	2.457	0.009	1.000

P-value adjusted for comparing a family of 6.

TABLE 9 Results of *post hoc* comparisons of analysis of variance (ANOVA) comparing the performance of different school grades (4th to 9th) on the FL subtest of AREF.

Grade		Mean difference	SE	*t*	p_{tukey}
4	5	−0.058	1.454	−0.040	1.000
	6	−3.532	1.651	−2.139	0.277
	7	−5.172	1.623	−3.186	0.024
	8	−5.404	1.717	−3.147	0.027
	9	−6.639	2.213	−3.000	0.040
5	6	−3.475	1.511	−2.299	0.206
	7	−5.115	1.481	−3.454	0.011
	8	−5.347	1.583	−3.377	0.014
	9	−6.582	2.111	−3.118	0.029
6	7	−1.640	1.675	−0.979	0.923
	8	−1.872	1.766	−1.060	0.896
	9	−3.107	2.251	−1.380	0.739
7	8	−0.232	1.740	−0.133	1.000
	9	−1.467	2.231	−0.657	0.986
8	9	−1.235	2.300	−0.537	0.994

P-value adjusted for comparing a family of 6.

TABLE 10 Results of *post hoc* comparisons of analysis of variance (ANOVA) comparing the performance of different school grades (4th to 9th) on the WM subtest of AREF.

Grade		Mean difference	SE	*t*	p_{tukey}
4	5	−1.152	2.171	−0.531	0.995
	6	−8.749	2.465	−3.549	0.008
	7	−11.570	2.424	−4.773	<0.001
	8	−11.575	2.564	−4.514	<0.001
	9	−12.716	3.305	−3.848	0.003
5	6	−7.597	2.256	−3.367	0.014
	7	−10.418	2.211	−4.711	<0.001
	8	−10.423	2.364	−4.409	<0.001
	9	−11.564	3.152	−3.669	0.005
6	7	−2.821	2.501	−1.128	0.869
	8	−2.826	2.637	−1.071	0.891
	9	−3.967	3.362	−1.180	0.845
7	8	−0.005	2.599	−0.002	1.000
	9	−1.146	3.332	−0.344	0.999
8	9	−1.141	3.435	−0.332	0.999

P-value adjusted for comparing a family of 6.

be significant and positive. This finding reinforces the construct validity of the instrument, considering that the FDT demonstrates correlations with inhibitory control measures (De Paula et al., 2017). Previous studies also corroborated a higher correlation between reading comprehension tasks and Verbal IQ compared to Performance IQ (López-Escribano et al., 2013; Ready et al., 2013).

Similarly to those AREF subtests, the Flexibility subtest demonstrated positive correlations with PROLEC and PROLEC-SE-R results, with the inhibition percentile of the FDT and the verbal IQ measure as well. However, like the GSF task, the Flexibility subtest showed correlation only with the inhibitory control measure, not demonstrating correlation with the

TABLE 11 Comparative analysis using student's *t*-test of performance in GSF, IC-2 and WM subtests between students from private and public schools in grades 6th to 9th.

Subtest	*t*	df	*p*	Cohen's d	SE Cohen's d
GSF efficiency	−0.792	48	0.433	−0.233	0.297
IC-2 efficiency	−1.477	48	0.146	−0.435	0.303
Total WM	−2.135	48	0.038*	−0.629	0.313

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. GSF Efficiency, Efficiency of graphophonological-semantic flexibility subtest. IC-2 Efficiency, Efficiency of the second task of inhibitory control subtest. Total WM, Total points obtained in the 7 items of the AREF working memory task.

TABLE 12 Means and standard deviations for public and private school participants (6th to 9th grade) on GSF, IC-2, and WM subtests.

Subtest	School type	*N*	Mean	Standard deviation
GSF Efficiency	Public	18	2.748	1.314
	Private	32	3.069	1.414
IC-2 efficiency	Public	18	10.646	4.862
	Private	32	13.043	5.830
Total WM	Public	18	25.500	6.510
	Private	32	29.375	5.961

GSF Efficiency, efficiency of graphophonological-semantic flexibility subtest. IC-2 Efficiency, Efficiency of the second task of inhibitory control subtest. Total WM, Total points obtained in the 7 items of the AREF working memory task.

TABLE 13 Comparison between public and private schools with participants from 6th to 9th grade - Mann–Whitney test for IC-1 and FL subtests.

Subtest	*w*	*p*	Rank-Biserial correlation
IC-1 efficiency	187.500	0.043*	−0.349
FL efficiency	190.000	0.049*	−0.340

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. IC-1 Efficiency, Efficiency of the first task of inhibitory control subtest. FL Efficiency, Efficiency of the flexibility subtest.

TABLE 14 Medians and standard deviations for public and private school participants (6th to 9th grade) on IC-1 and FL subtests.

Subtest	School type	*N*	Median	Standard deviation
IC-1 Efficiency	Public	18	11.325	4.192
	Private	32	13.605	3.601
FL Efficiency	Public	18	9.575	4.841
	Private	32	12.700	5.381

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. IC-1 Efficiency, Efficiency of the first task of inhibitory control subtest. FL Efficiency, Efficiency of the flexibility subtest.

flexibility measure. One possible explanation for this finding is that, despite the task being initially developed to measure flexibility, it may recruit more inhibition processes. It is noteworthy, however, the previously mentioned observation that inhibition precedes flexibility (Diamond, 2013). Therefore, these initial findings may suggest the recruitment of this process in the task, something that should be investigated in future studies with larger samples.

TABLE 15 Reliability statistics of the graphophonological-semantic flexibility (GSF) subtest.

Estimate	Cronbach's α
Point estimate	0.566
95% CI lower bound	0.502
95% CI upper bound	0.630

CI, Confidence interval.

TABLE 16 Reliability statistics of the items in the graphophonological-semantic flexibility (GSF) subtest.

Item	Item-rest correlation
Matrix 1 - Points	−0.231
Matrix 1 – Justification	−0.240
Matrix 1 – Time	0.486
Matrix 2 – Points	−0.295
Matrix 2 – Justification	−0.095
Matrix 2 – Time	0.740
Matrix 3 – Points	−0.266
Matrix 3 – Justification	−0.185
Matrix 3 - Time	0.630

TABLE 17 Reliability statistics of the Inhibitory Control and Flexibility subtests.

Estimate	Cronbach's α
Point estimate	0.768
95% CI lower bound	0.753
95% CI upper bound	0.786

CI, Confidence interval.

TABLE 18 Reliability statistics of the items in the Inhibitory Control and Flexibility subtests.

Item	Item-rest correlation
BT - Reading time	0.924
BT - Retelling	−0.361
BT - Questions	−0.290
IC-1 - Reading time	0.945
IC-1 - Retelling	−0.161
IC-1 - Questions	−0.165
IC-2 - Reading time	0.942
IC-2 - Retelling	−0.077
IC-2 - Questions	−0.344
FL - Reading time	0.777
FL - Retelling	−0.075
FL - Questions	−0.074

TABLE 19 Reliability statistics of the working memory subtest.

Estimate	Cronbach's α
Point estimate	0.881
95% CI lower bound	0.848
95% CI upper bound	0.909

CI, Confidence interval.

TABLE 20 Reliability statistics of the items in the working memory subtest.

Item	Item-rest correlation
WM1	0.493
WM2	0.678
WM3	0.704
WM4	0.661
WM5	0.724
WM6	0.738
WM7	0.840

In relation to inhibitory control tasks, our initial hypothesis was that an increase in distractors would lead to reduced reading efficiency compared to the performance observed in the Baseline Text. Indeed, the study by Borella and De Ribaupierre (2014) identified that resistance to distractors, measured through an external task to reading assessment (Color Stroop Task), was one of the predictors of text comprehension. However, the analyses conducted with the two inhibitory control tasks comprising the AREF resemble more closely those conducted in studies where distractors were part of the text read by participants (Connelly et al., 1991; Kemper and McDowd, 2006), and the presence of these elements was associated with reduced reader performance. Similarly to these previous studies, in the current research, participant performance was significantly lower when distractors were present in the text, albeit this was observed only in the second inhibitory control task (IC-2). In the context of the AREF tasks, the significant difference observed in the comparison between Baseline Text and IC-2 can be explained by the presence of two distractor stimuli (colored lines and target color words that should not be read) in the latter. The inclusion of more distractors may have increased the cognitive demand of the task, possibly resulting in lower average performance. Performance on the Flexibility task did not show significant differences compared to the Baseline Text. Although previous studies have shown unique contributions of flexibility to reading comprehension (Colé et al., 2014; Hung and Loh, 2021), as far as we know, no research has investigated cognitive flexibility during the reading of a text that required response alternation, such as the AREF task. Our initial hypothesis was that the demand of the flexibility task would reduce its efficiency, but this hypothesis was not confirmed. Therefore, further investigation with a larger sample is needed to confirm the consistency of the results of the inhibitory control and flexibility tasks.

In the Working Memory (WM) subtest, significant, positive and moderate correlations were observed between task results and external measures, such as reading subtests from PROLEC and PROLEC-SE-R, digit span in forward order, Scaled Scores of the WISC-IV Digit Span task, and the T-score of the Vocabulary subtest of the WASI. It is noteworthy that the latter correlation proved to be more robust than that observed between the AREF subtests and the T-score of the WASI Matrix Reasoning subtest, as predicted in the hypotheses formulated. These results not only provide support for external construct validity but also corroborate previous conclusions. For example, this is consistent with evidence that vocabulary and working memory are predictive of reading performance in children, as highlighted by Piccolo and Salles (2013). Another study (Babayiğit, 2015) indicated that differences in reading comprehension performance between individuals who had English as their first language (L1) and those who had English as their second language (L2) were explained by differences in oral language skills in English (including vocabulary and verbal working memory), with higher scores in the L1 group in both textual comprehension and oral language skills. Longitudinal data (Holahan et al., 2018), following students from grades 1 to 9, also found unique contributions of vocabulary to the development of reading comprehension. Additionally, as emphasized in the review by Butterfuss and Kendeou (2018), working memory plays an essential role in reading comprehension, as the central executive component facilitates restricting information in the phonological loop, especially in contexts where sentences become longer and syntactically more complex. This observation is consistent with the results of this study, in which the Working Memory task demands greater use of working memory as texts become longer. Nevertheless, the correlation between the AREF Working Memory result and backward digit span did not reach significance, contrary to our initial hypothesis.

The data from the present study indicated variations in the performance of the AREF subtests among participants from different school grades and between those from public and private schools. These results support the concurrent validity of the tool.

First, it was found that, in all subtests, there were statistically significant differences in student performance, with the 4th and 5th-grade results being notably lower than those of other grades in most comparisons. These findings are consistent with developmental literature, which reports cognitive improvements in the age range covered by this study, both in terms of executive functions (Jacobsen et al., 2017) and reading comprehension (De Oliveira et al., 2023). It should be noted, however, that this effect may also be associated with the presence of only public school students in the 4th and 5th-grade sample. Future studies should include younger students from private schools to verify if this result remains robust.

To prevent the average performance of public school participants from being lowered due to the inclusion of younger grades, the analyses comparing the performance of individuals from public and private schools on AREF subtests were conducted only with students from the 6th to 9th grades, as these groups included students from both types of schools.The results revealed higher average scores among private school students in WM, IC-1, and FL tasks, with no significant differences in GSF and IC-2 tasks. Although the present study did not collect data on participants' socioeconomic status, the differences between the two groups may be related to this factor, as found in other studies (Cáceres-Serrano and Alvarado-Izquierdo, 2017; Çigdemir and Akyol, 2022; Jacobsen et al., 2017).

This study also presented indications of reliability of the AREF instrument. Regarding internal consistency, the Graphophonological-Semantic Flexibility subtest showed low levels of consistency, possibly due to their multidimensionality and the sample size (Cortina, 1993). The hypothesis of multidimensionality can be raised because the

items comprising that subtest involve both scores related to the correct task performance and time measures.

On the other hand, in relation to Inhibitory Control and Flexibility, Cronbach's alpha coefficient indicated moderate internal consistency, while the item-total correlation revealed that performance on specific items correlated weakly to strongly with total task performance. Notably, the most strongly correlated items with overall task performance were those related to timing measures, indicating that shorter reading periods were associated with better performance in the AREF. The same result was observed in the Graphophonological-Semantic Flexibility subtest, where the score on the scale was negatively related to the time spent on its completion. These observations are in line with evidence suggesting a negative relationship between accuracy in executive function tests and execution time (Camerota et al., 2019). Similarly, reduced reading speed is related to overload in working memory, resulting in reduced availability of attentional resources for reading comprehension (Zoccolotti et al., 2016; Rispens, 2004). Therefore, regarding the assessment of the two main constructs measured by the AREF - Reading Comprehension and Executive Functions -, the data suggest that longer task completion times are associated with inferior performance, which was supported by this study.

In contrast, the internal consistency of the Working Memory task was considered high, with items showing strong positive correlations with overall task performance. It indicates good reliability of this task.

The results of this study corroborate previous findings highlighting the interdependence of executive functions, such as inhibitory control, cognitive flexibility and working memory, with reading skills. However, it is crucial to interpret these results in light of the study's limitations. Firstly, it is important to note that the research did not include a sample from the private school population of 4th and 5th grade elementary school students. Another relevant limitation is the composition of the recruited participants. Although there was a variety of age ranges, covering students aged 8 to 14 years, the study had a relatively small sample of students. Additionally, the research focused exclusively on students from the southeastern region of Brazil, which may limit the generalization of the results to the overall population. Lastly, another limitation of this study is the lack of socioeconomic data that could have been included in the statistical analyses. The inclusion of these data could be important for interpreting the results, especially considering that socioeconomic factors have shown significant correlations with both vocabulary and reading comprehension development (Lervåg et al., 2019; Olsen and Huang, 2022) as well as executive functions (Last et al., 2018; Lawson et al., 2018).

Based on the results obtained, it is possible to conclude that the AREF instrument presents initial psychometric evidence indicating its viability for clinical and research use after obtaining a normative sample. Although the strengths of the correlations with other instruments range from weak to moderate, this can be attributed to the many factors influencing performance on the complex target constructs: reading comprehension and executive functions. Considering the complexity of evaluating both constructs, it is a significant achievement that the test has demonstrated construct validity evidence for both variables, indicating its utility, especially in the Brazilian context, where no equivalent exists.

However, it is evident that further studies are necessary to reinforce the psychometric validation of the developed subtests. Specifically, the lack of correlation of the Flexibility task and Graphophonological-Semantic Flexibility of the AREF with external measures of flexibility highlights the need for a more in-depth investigation to determine if the subtests are truly assessing what it intends to. Additionally, for the IC/FL and GSF subtests, it would be important to conduct further reliability analysis using methods more sensitive to the multidimensionality of the tasks. Without these analyses, the scores obtained by individuals undergoing the application should be interpreted with caution. Regarding the working memory subtest, where time is not a variable, the measure of external consistency was high, and the correlations with external measures support its construct validity, suggesting it is suitable for use.

Future studies with larger and more representative samples are essential to replicate the findings obtained and determine if these findings can be extrapolated to other populations. Additionally, it is crucial to conduct further research to evaluate the instrument's sensitivity regarding students reporting difficulties in reading comprehension and executive deficits, both in the presence and absence of mental disorders. Furthermore, the importance of conducting normative studies to establish parameters that allow for the interpretation of data obtained with students and patients is emphasized.

Despite significant challenges associated with creating tasks capable of simultaneously assessing reading processes and executive functions, the findings of this study suggest that the AREF appears to fulfill this complex purpose effectively. This finding has promising implications, indicating that the AREF may be a useful tool in the neuropsychological assessment of children and adolescents with reading comprehension difficulties, as well as in cases of isolated executive dysfunctions or as part of various neurodevelopmental disorders, including specific learning disorders and attention deficit hyperactivity disorder (ADHD).

Furthermore, the data obtained through the AREF have the potential to support the planning of therapeutic interventions in various areas, including neuropsychology, speech therapy, and educational psychology. A deeper understanding of the performance patterns of these individuals will allow for a more personalized approach to help them overcome their specific difficulties.

Data availability statement

The original statistical analysis presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Comitê de Ética em Pesquisa da Universidade Federal de Minas Gerais – COEP-UFMG (Ethics Committee in Research of the Universidade Federal de Minas Gerais). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

VO: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. JV-M: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. AP: Writing – review & editing, Formal analysis. RF: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. LM-D: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. We acknowledge CAPES, CNPq, and PRPq. The authors are affiliated with the National Institutes of Science and Technology Program - INCT-Neurotec_R. Grant #406935/2022-0.

Acknowledgments

We extend our sincere gratitude to Fernanda Luísa, Gabriella Vilaça, Gustavo Fretta, Karla Nietsche, Laura Ludgero, Luana Lobo, Maria Júlia Veloso, Rafaela Guatimosim, Rebeca Rodarte, Sophia Lima, Silvia Assis, Yumi de Halley, and Yuri Banov for their invaluable contributions, which made AREF (ALEFE) possible.

Conflict of interest

The AREF (ALEFE) is in negotiations for potential commercialization by the publisher AMPLA, which could, in the future, remunerate the researchers.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1399388/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Comparison of the average performance of individuals from public and private schools on the WM task.

SUPPLEMENTARY FIGURE S2
Comparison of the average performance of individuals from public and private schools on the IC-1 task.

SUPPLEMENTARY FIGURE S3
Comparison of the average performance of individuals from public and private schools on the FL task.

References

American Educational Research Association (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing.

Babayiğit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): a multigroup structural analysis. *Read. Writ.* 28, 527–544. doi: 10.1007/s11145-014-9536-x

Best, J. R., Miller, P. H., and Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learn. Individ. Differ.* 21, 327–336. doi: 10.1016/j.lindif.2011.01.007

Borella, E., and De Ribaupierre, A. (2014). The role of working memory, inhibition, and processing speed in text comprehension in children. *Learn. Individ. Differ.* 34, 86–92. doi: 10.1016/j.lindif.2014.05.001

Burgess, A. N., and Cutting, L. E. (2023). The Behavioral and neurobiological relationships between executive function and Reading: a review of current and preliminary findings. *Mind Brain Educ.* 17, 267–278. doi: 10.1111/mbe.12378

Butterfuss, R., and Kendeou, P. (2018). The role of executive functions in reading comprehension. *Educ. Psychol. Rev.* 30, 801–826. doi: 10.1007/s10648-017-9422-6

Cáceres-Serrano, P., and Alvarado-Izquierdo, J. M. (2017). The effect of contextual and socioeconomic factors on reading comprehension levels. *Mod. J. Lang. Teach. Methods* 7, 76–85.

Camerota, M., Willoughby, M. T., and Blair, C. B. (2019). Speed and accuracy on the hearts and flowers task interact to predict child outcomes. *Psychol. Assess.* 31, 995–1005. doi: 10.1037/pas0000725

Capellini, S. A., Oliveira, A. M., and Cuetos, F. (2012). PROLEC: Provas de Avaliação dos Processos de Leitura. *2nd* Edn. São Paulo, SP: Casa do Psicólogo.

Capilla, A., Romero, D., Maestú, F., Campo, P., Fernández, S., González-Marqués, J., et al. (2004). Emergence and brain development of executive functions. *Actas Esp. Psiquiatr.* 32:377.

Cartwright, K. B. (2007). The contribution of Graphophonological-semantic flexibility to Reading comprehension in college students: implications for a less simple view of Reading. *J. Lit. Res.* 39, 173–193. doi: 10.1080/10862960701331902

Cartwright, K. B., Marshall, T. R., Dandy, K. L., and Isaac, M. C. (2010). The development of graphophonological-semantic cognitive flexibility and its contribution to reading comprehension in beginning readers. *J. Cogn. Dev.* 11, 61–85. doi: 10.1080/15248370903453584

Cartwright, K. B., Marshall, T. R., Huemer, C. M., and Payne, J. B. (2019). Executive function in the classroom: cognitive flexibility supports reading fluency for typical readers and teacher-identified low-achieving readers. *Res. Dev. Disabil.* 88, 42–52. doi: 10.1016/j.ridd.2019.01.011

Çigdemir, S., and Akyol, H. (2022). The relationship between environmental factors and Reading comprehension. *Int. J. Progress. Educ.* 18, 150–164. doi: 10.29329/ijpe.2022.439.11

Colé, P., Duncan, L. G., and Blaye, A. (2014). Cognitive flexibility predicts early reading skills. *Front. Psychol.* 5:565. doi: 10.3389/fpsyg.2014.00565

Collins, A. A., and Lindström, E. R. (2021). Making sense of reading comprehension assessments: guidance for evaluating student performance. *Interv. Sch. Clin.* 57, 23–31. doi: 10.1177/1053451221994806

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Connelly, S. L., Hasher, L., and Zacks, R. T. (1991). Age and reading: the impact of distraction. *Psychol. Aging* 6, 533–541. doi: 10.1037/0882-7974.6.4.533

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *J. Appl. Psychol.* 78, 98–104. doi: 10.1037/0021-9010.78.1.98

Cuetos, F., Arribas, D., and Ramos, J. R. (2022). PROLEC-SE-R: Provas de Avaliação dos Processos de Leitura - Ensino Fundamental II e Médio. São Paulo: Hogrefe.

Dehaene, S. (2009). Reading in the brain: The science and evolution of a human invention. New York: Viking.

De Oliveira, A. M., Santos, J. L. F., and Capellini, S. A. (2023). Reading comprehension performance of elementary and senior high school students. *Front. Educ.* 8:1086040. doi: 10.3389/feduc.2023.1086040

De Paula, J. J., Oliveira, T. D., Querino, E. H. G., and Malloy-Diniz, L. F. (2017). The five digits test in the assessment of older adults with low formal education: construct validity and reliability in a Brazilian clinical sample. *Trends Psychiatry Psychother.* 39, 173–179. doi: 10.1590/2237-6089-2016-0060

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Follmer, D. J. (2018). Executive function and reading comprehension: a meta-analytic review. *Educ. Psychol.* 53, 42–60. doi: 10.1080/00461520.2017.1309295

Fonseca, R. P., Seabra, A. G., and Miranda, M. C. (2020). "Neuropsicologia escolar: revisitando conceitos e práticas" in Neuropsicologia Escolar. eds. R. P. Fonseca, A. G. Seabra and M. C. Miranda (São Paulo: Pearson Clínica Brasil), 55–95.

Gonçalves, H. A., Viapiana, V. F., Sartori, M. S., Giacomoni, C. H., Stein, L. M., and Fonseca, R. P. (2017). Funções executivas predizem o processamento de habilidades básicas de leitura, escrita e matemática? *Neuropsicol. Latinoam.* 9, 42–54. doi: 10.5579/rnl.2016.039

Gough, P. B., and Tunmer, W. E. (1986). Decoding, Reading, and Reading disability. *Remedial Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Holahan, J. M., Ferrer, E., Shaywitz, B. A., Rock, D. A., Kirsch, I. S., Yamamoto, K., et al. (2018). Growth in reading comprehension and verbal ability from grades 1 through 9. *J. Psychoeduc. Assess.* 36, 307–321. doi: 10.1177/0734282916680984

Huizinga, M., Dolan, C. V., and Van der Molen, M. W. (2006). Age-related change in executive function: developmental trends and a latent variable analysis. *Neuropsychologia* 44, 2017–2036. doi: 10.1016/j.neuropsychologia.2006.01.010

Hung, C. O. Y., and Loh, E. K. Y. (2021). Examining the contribution of cognitive flexibility to metalinguistic skills and reading comprehension. *Educ. Psychol.* 41, 712–729. doi: 10.1080/01443410.2020.1734187

Jacobsen, G. M., de Mello, C. M., Kochhann, R., and Fonseca, R. P. (2017). Executive functions in school-age children: influence of age, gender, school type and parental education. *Appl. Cogn. Psychol.* 31, 404–413. doi: 10.1002/acp.3338

JASP Team. (2023). *JASP (version 0.17.2.0) [computer software]*. Available at: https://jasp-stats.org/.

Kemper, S., and McDowd, J. (2006). Eye movements of young and older adults while reading with distraction. *Psychol. Aging* 21:32:39. doi: 10.1037/0882-7974.21.1.32

Kendeou, P., Van Den Broek, P., Helder, A., and Karlsson, J. (2014). A cognitive view of reading comprehension: implications for reading difficulties. *Learn. Disabil. Res. Pract.* 29, 10–16. doi: 10.1111/ldrp.12025

Kieffer, M. J., Vukovic, R. K., and Berry, D. (2013). Roles of attention shifting and inhibitory control in fourth-grade reading comprehension. *Read. Res. Q.* 48, 333–348. doi: 10.1002/rrq.54

Last, B. S., Lawson, G. M., Breiner, K., Steinberg, L., and Farah, M. J. (2018). Childhood socioeconomic status and executive function in childhood and beyond. *PLoS One* 13:e0202964. doi: 10.1371/journal.pone.0202964

Latzman, R. D., Elkovitch, N., Young, J., and Clark, L. A. (2010). The contribution of executive functioning to academic achievement among male adolescents. *J. Clin. Exp. Neuropsychol.* 32, 455–462. doi: 10.1080/13803390903164363

Lawson, G. M., Hook, C. J., and Farah, M. J. (2018). A meta-analysis of the relationship between socioeconomic status and executive function performance among children. *Dev. Sci.* 21:e12529. doi: 10.1111/desc.12529

Lervåg, A., Dolean, D., Tincas, I., and Melby-Lervåg, M. (2019). Socioeconomic background, nonverbal IQ and school absence affects the development of vocabulary and reading comprehension in children living in severe poverty. *Dev. Sci.* 22:e12858. doi: 10.1111/desc.12858

López-Escribano, C., De Juan, M. R. E., Gómez-Veiga, I., and García-Madruga, J. A. (2013). A predictive study of reading comprehension in third-grade Spanish students. *Psicothema* 25, 199–205. doi: 10.7334/psicothema2012.175

Magnus, B. E., Willoughby, M. T., Blair, C. B., and Kuhn, L. J. (2019). Integrating item accuracy and reaction time to improve the measurement of inhibitory control abilities in early childhood. *Assessment* 26, 1296–1306. doi: 10.1177/1073191117740953

Marques de Oliveira, A., Santos, J. L. F., and Capellini, S. A. (2024). Reading processes of public and private middle school and high school students. *Psicol. Reflex. Crít.* 37:14. doi: 10.1186/s41155-024-00296-0

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Nyongesa, M. K., Ssewanyana, D., Mutua, A. M., Chongwo, E., Scerif, G., Newton, C. R., et al. (2019). Assessing executive function in adolescence: a scoping review of existing measures and their psychometric robustness. *Front. Psychol.* 10:311. doi: 10.3389/fpsyg.2019.00311

OECD (2023). PISA 2022 results (volume I): the state of learning and equity in education. Paris: PISA, OECD Publishing.

Olsen, A. A., and Huang, F. L. (2022). Interaction of socioeconomic status and class relations on reading. *J. Lit. Res.* 54, 346–369. doi: 10.1177/1086296X2211168

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., et al. (2018). A meta-analysis on the relation between reading and working memory. *Psychol. Bull.* 144:48:76. doi: 10.1037/bul0000124

Peng, P., and Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: a bidirectional perspective. *Child Dev. Perspect.* 14, 15–20. doi: 10.1111/cdep.12352

Perfetti, C. A., and Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *J. Educ. Psychol.* 67, 461–469. doi: 10.1037/h0077013

Rabiner, D. L., Godwin, J., and Dodge, K. A. (2016). Predicting academic achievement and attainment: the contribution of early academic skills, attention difficulties, and social competence. *Sch. Psychol. Rev.* 45, 250–267. doi: 10.17105/SPR45-2.250-267

Piccolo, L. D. R., and Salles, J. F. (2013). Vocabulary and working memory predict reading performance of children. *Psicol. Teor. Prát.* 15, 180–191.

Ready, R. E., Chaudhry, M. F., Schatz, K. C., and Strazzullo, S. (2013). "Passageless" Administration of the Nelson–Denny Reading Comprehension Test: associations with IQ and Reading skills. *J. Learn. Disabil.* 46, 377–384. doi: 10.1177/0022219412468160

Rispens, J. E. (2004). Syntactic and phonological processing in developmental dyslexia. Groningen: Rijksuniversiteit Groningen, Faculteit der Letteren.

Romine, C. B., and Reynolds, C. R. (2005). A model of the development of frontal lobe functioning: findings from a meta-analysis. *Appl. Neuropsychol.* 12, 190–201. doi: 10.1207/s15324826an1204_2

Sedó, M., De Paula, J. J., and Maloy-Diniz, L. F. (2015). O Teste dos 5 Dígitos. São Paulo: Hografe.

Spencer, M., Richmond, M. C., and Cutting, L. E. (2020). Considering the role of executive function in reading comprehension: a structural equation modeling approach. *Sci. Stud. Read.* 24, 179–199. doi: 10.1080/10888438.2019.1643868

Su, S., and Davison, M. L. (2019). Improving the predictive validity of reading comprehension using response times of correct item responses. *Appl. Meas. Educ.* 32, 166–182. doi: 10.1080/08957347.2019.1577247

Varghese, S. M., and Shanbal, J. C. (2024). Profiling of Graphophonological semantic flexibility in typical readers: a cross-sectional study. *Indian J. Psychol. Med.* 2024:1252. doi: 10.1177/02537176241252

Verhoeven, L., and Perfetti, C. A. (2011). Morphological processing in reading acquisition: a cross-linguistic perspective. *Appl. Psycholinguist.* 32, 457–466. doi: 10.1017/S0142716411000154

Wechsler, D. (2013). "Escala Wechsler de Inteligência para Crianças de Adolescentes – 4ª" in Edição (WISC-IV). Manual para Administração e Avaliação (São Paulo: Casa do Psicólogo).

Wechsler, D. (2014). Escala Abreviada de inteligência Wechsler. Primeira Edição. São Paulo: Casa do Psicólogo.

Zoccolotti, P., De Jong, P. F., and Spinelli, D. (2016). Understanding developmental dyslexia: linking perceptual and cognitive deficits to reading processes. *Front. Hum. Neurosci.* 10:140. doi: 10.3389/fnhum.2016.00140

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 19 November 2024
DOI 10.3389/fpsyg.2024.1448817

OPEN ACCESS

EDITED BY
Elisa Cavicchiolo,
University of Rome Tor Vergata, Italy

REVIEWED BY
Paolo Antonino Grasso,
University of Florence, Italy
Andrew Anderson,
The University of Melbourne, Australia

*CORRESPONDENCE
Giulia Carlotta Rizzo
✉ giuliacarlotta.rizzo@unimib.it

†These authors have contributed equally to this work and share first authorship

RECEIVED 13 June 2024
ACCEPTED 28 October 2024
PUBLISHED 19 November 2024

CITATION
De Luca M, Nardo D, Rizzo GC, Daini R, Tavazzi S and Zeri F (2024) Short Italian Wilkins Rate of Reading Test for repeated-measures designs in optometry and neuropsychology. *Front. Psychol.* 15:1448817. doi: 10.3389/fpsyg.2024.1448817

Short Italian Wilkins Rate of Reading Test for repeated-measures designs in optometry and neuropsychology

Maria De Luca [1†], Davide Nardo [2†], Giulia Carlotta Rizzo [3,4*], Roberta Daini [4,5], Silvia Tavazzi [3,4] and Fabrizio Zeri [3,4]

[1]IRCCS Fondazione Santa Lucia, Rome, Italy, [2]Department of Education, University of Roma Tre, Rome, Italy, [3]Department of Materials Science, University of Milano-Bicocca, Milan, Italy, [4]COMiB Research Centre in Optics and Optometry, University of Milano-Bicocca, Milan, Italy, [5]Department of Psychology, University of Milano-Bicocca, Milan, Italy

Background: The recently published New Italian version of the Wilkins Rate of Reading Test (standard Italian WRRT) was designed to measure reading speed in repeated-measures designs in research and/or clinical examinations. The test features 15 equivalent 10-line passages made up of unrelated words, adhering to the principles established by the Wilkins Rate of Reading Test in English (original WRRT).

Aim: To develop a short Italian version of the WRRT (SI-WRRT), and to determine the equivalence across the new, shorter passages of text. The introduction of 5-line passages, instead of the original 10-line ones, aims to enhance the tool's suitability for the elderly or neuropsychological patients by reducing administration time.

Method: The same 15 high-frequency Italian words from the standard Italian WRRT were used to generate 15 5-line passages for the SI-WRRT. Comprehensive eye examination and vision assessment, including the Radner Reading Charts, were performed before the administration of the SI-WRRT. Forty healthy Italian-speaking higher education students read the SI-WRRT passages aloud in random order. Reading speed and accuracy were measured offline from digital recordings of the readings. Equivalence across passages and the effects of practice and fatigue were assessed for reading speed and accuracy, along with test-retest reliability.

Results: No significant difference in reading speed was found across 14 out of the 15 passages. In addition, no differences were observed in accuracy, and the error rate was very low. Practice and fatigue effects were minimal for reading speed, whereas they were absent for accuracy. Reading speed, the reference metric for the WRRT, showed moderate-to-good test-retest reliability.

Conclusions: Equivalence was confirmed across 14 passages of the SI-WRRT. Therefore, the test may be suitable for examining the elderly or neuropsychological patients, as reading time of the 5-line passages is halved with respect to the standard Italian WRRT. However, the 5-line passages still allow the assessment of prolonged reading. Since one passage was not equivalent, we recommend avoiding the use of random rearrangements of words without formally checking their validity.

KEYWORDS

Wilkins Rate of Reading Test, wpm, reading speed, repeated-measures, equivalent texts, Radner Reading Charts, practice effect, fatigue effect

1 Introduction

In scientific research, there is often a need to measure the dependent variable more than once. In such situations, repeated-measures designs are commonly adopted, wherein the same participants are enrolled in experimental sessions in which a dependent variable is measured on multiple occasions over time (e.g., pre-, post-, and follow-up testing), or under different experimental conditions. For instance, repeated-measures designs are adopted when the same group of participants is exposed to different interventions or to an intervention protocol vs. a control condition. Whenever a measure is repeated, there is a potential issue with the equivalence of alternative test versions (e.g., Beglinger et al., 2005).

In the field of vision science, which encompasses disciplines ranging from optometry and ophthalmology to cognitive psychology and neuropsychology, repeated-measures designs based on multiple readings of texts (over time or across conditions) are typically applied to assess the efficacy of interventions for patients suffering from various vision deficits (e.g., Bailey and Lakshminaryanan, 1997), or reading interventions for patients with developmental (i.e., dyslexia; e.g., Tilanus et al., 2019; autism spectrum disorders; e.g., Ludlow et al., 2006) or acquired reading deficits (e.g., hemianopic alexia after stroke or traumatic brain injury; e.g., Spitzyna et al., 2007).

Irrespective of whether texts are used as diagnostic/monitoring tools in a clinical context, or as stimulus materials in experimental designs (e.g., Wilkins et al., 2005; Zeri et al., 2018), the same passage should never be used for multiple readings, to avoid practice/learning effects. Therefore, different passages need to be used, provided that they represent parallel forms (i.e., equivalent texts) that do not introduce factors that may interfere with the manipulated variable(s) and may hence produce unreliable (or hard-to-interpret) results. Different passages are equivalent when their basic characteristics (e.g., total number of words, syllables, and characters; number of words per line; number of lines of text) and text complexity (e.g., word frequency; syntax; sentence length; clause complexity; semantics) are comparable (e.g., Brussee et al., 2015; Radner et al., 2017; Trauzettel-Klosinski et al., 2012).

An alternative, effective way to generate homogeneous material for serial readings is to minimise the linguistic content of a text by using unrelated, high frequency words arranged in random order within a text line (e.g., Bailey and Lovie, 1980; Wilkins et al., 1996). This way, reading relies only on basic reading skills (as in primary school), without requiring any higher cognitive processing (e.g., inferring a meaning to generate predictions). Then, reading becomes dependent only on single-word processing and on visuoperceptual features of the stimulus, without any contextual influence (e.g., Stanovich et al., 1985). This has the further advantage of making the material suitable for testing children and adults with modest linguistic skills, as done in the Wilkins Rate of Reading Test (from here on, "original WRRT;" Wilkins et al., 1996), which uses passages made up of unrelated, short, high-frequency words (i.e., passages which are meaningless at the sentence-level).

The original WRRT was designed and adopted in optometry and vision science to assess visual performance in reading under different visual conditions (e.g., the use of different coloured overlays to aid reading difficulties; Wilkins et al., 1996). The test comprises 10 lines of text containing the same 15 words (repeated line by line), which are very common in the English lexicon, arranged in random order. The rationale of this test is to return reading speed as "words correctly read per minute" (wpm) using reading materials that neutralise/minimise the impact of syntax and semantics on the task. That is, the text content is as simple as possible, does not convey any meaning at the sentence-level, and is matched across conditions, so that any effect can be solely attributed to the experimental manipulation or clinical intervention at hand, and not to the text itself. Since Wilkins et al.'s aim was also to create materials that elicit visual stress, words within each line were closely spaced and line spacing was tighter. This way, reading is visually—but not cognitively—demanding, allowing the investigation of the effect of visuoperceptual factors and interventions on reading (Evans and Joseph, 2002; Monger et al., 2015; Northway, 2003; Wilkins, 2002). In addition, "single passage" versions in other languages were made available on Wilkins' website (http://www1.essex.ac.uk/psychology/overlays/rrt%20OC4.htm), including a 20-line Italian passage, although their validity was not determined. To create more passages, Wilkins et al. suggested generating equivalent forms of texts by randomly rearranging the words within each line. Therefore, multiple versions (up to four passages) of the original test were made available (Wilkins et al., 1996; Gilchrist et al., 2021). However, such passages were only assessed in terms of test-retest reliability (Gilchrist et al., 2021), rather than equivalence.

Conversely, we hypothesised that random rearrangements of high-frequency words *per se* may not necessarily return equivalent passages, as a given random order may accidentally form meaningful word sequences (which would otherwise be unrelated), possibly impacting on reading speed [e.g., "come see the play," or "you see the dog" in Wilkins et al. (1996)]. Such concern prompted us to assess the equivalence across passages, which was formally tested and confirmed (alongside test-retest reliability) in a recent study introducing the New Italian version of the Wilkins Rate of Reading Test (from here on "standard Italian WRRT;" Zeri et al., 2023). In that study, we also increased the number of passages for protocols requiring more than 4 experimental conditions or repeated measures. Hence, the standard Italian WRRT features 15 equivalent passages. The structure and constraints of the original WRRT were maintained, and a transliteration instead of a direct translation was adopted. In the standard Italian WRRT, participants achieved an average reading speed of 167.3 wpm. A passage was

Abbreviations: ANOVA, Analysis of Variance; arcsec, second of arc; BCVA, best corrected visual acuity; cd/m^2, candela per square metre; CPS, critical print size; D, dioptres; dpi, dots per inch; ICC, intraclass correlation coefficient; IreST, International Reading Speed Texts; logMAR, logarithm of the minimum angle of resolution (unit of measurement of print size); logRAD, logarithm of the Reading Acuity Determination (RAD), which is equivalent to the print size measured in logMAR adjusted for the reading errors made in the last sentence read entirely; Max, maximum; min, minute; Min, minimum; MNRead, Minnesota low vision reading chart; MRS, maximum reading speed; RA, reading acuity; RAN, rapid automatized naming; s, second; SD, standard deviation; SI-WRRT, Short Italian - Wilkins Rate of Reading Test; wpm, words per minute; WRRT, Wilkins Rate of Reading Test.

read in <1 min (mean ± SD: 54.9 ± 0.6 s, computed across mean values of single passages; range across individuals: 38.0–79.5). A session of 15 passages was completed in ~30 min (incl. 1 min of rest between passages).

A reading time of 55 s per text may represent a rather long duration in case of demanding tasks (e.g., any manipulation of passages display or layout that increases cognitive load and/or visual stress). In such cases, reading performance may be affected by fatigue, making it challenging to test multiple experimental conditions. A WRRT based on shorter equivalent passages would solve this issue. Shorter passages could also be helpful in studies involving the elderly or neuropsychological patients, who may present with attentional deficits, get quickly tired, or present with cognitive fatigability (e.g., Möller et al., 2014).

Therefore, the aim of the present study was to develop a Short Italian Wilkins Rate of Reading Test (SI-WRRT), a ready-to-print Italian version using 5 lines instead of the 10-line passages (as in the standard Italian WRRT, or in the original WRRT in English). While maintaining the same 15 high-frequency words and number of passages (15) of the standard Italian WRRT, the present study examined whether the 5-line passages retained the same characteristics of the 10-line ones in terms of equivalence, practice and fatigue effects, and test-retest reliability. Although the 5-line passages originated from the standard Italian WRRT (Zeri et al., 2023), the equivalence was re-assessed, because the smaller layout size of the 5-line text (and resulting shorter reading time) may introduce reading speed differences across passages.

2 Materials and methods

All procedures and the use of optometric tests and reading materials were undertaken in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University of Milano-Bicocca (Prot. N. 0398635 del 30/10/2023 – UOR: 003406).

2.1 Participants

Higher-education students from the University of Milano-Bicocca (Milan, Italy) were recruited. A thorough eye examination and vision assessment, based on a standard optometric examination and a standard assessment of near functional vision during reading, were carried out to include only participants capable of fluently reading at near (cf. inclusion criteria outlined in Table 1), to ensure a reliable assessment of equivalence across the 15 SI-WRRT passages. All volunteers provided written informed consent. G*Power software (www.gpower.hhu.de) was used to determine the sample size for a repeated-measures design (ANOVA and paired comparisons) and test-retest reliability. An effect size of 0.66 was calculated for a reading speed difference between two measurements of 10 wpm, which is a clinically relevant difference in optometry and ophthalmology (Altpeter et al., 2015; Kaltenegger et al., 2019; Stöhr et al., 2024), consistent with previous measures obtained from our laboratory (unpublished data). A difference of < 10 wpm indicates comparable reading speeds for different texts. Using a standard $\alpha = 0.05$ and a power $(1\text{-}\beta) = 0.80$ returned a minimum sample size of $n = 21$. However, in the present study we adopted a "conservative approach" and decided to double the sample size ($n = 42$) to increase statistical power, that is, the probability of correctly rejecting the null hypothesis (i.e., equivalence of passages), hence increasing the probability of identifying the presence of non-equivalent passages, if there are any. Based on this, 42 participants were enrolled. One participant had to be excluded due to the presence of developmental dyslexia identified during the medical history assessment. Another one dropped out of the study. There were no participants with visual profiles unsuitable for reading at near. Hence, our final sample included 40 participants (23 females and 17 males; mean age: 24.2 ± 3.7 years, range 19.0 – 35.0; mean years of education: 16.0 ± 2.0, range 13 – 21 years) all of whom returned for retesting after 2 weeks.

TABLE 1 Inclusion criteria for participants enrolled in the study (see Section 2.3).

Inclusion criteria
Native Italian speakers
Absence of known reading disability
No ocular pathology
No significant ocular motility or binocular vision anomalies (including strabismus)
Monocular best-corrected visual acuity (BCVA) at distance ≤ 0.10 logMAR in each eye
Near point of convergence ≤ 10 cm
Stereoscopic acuity ≤ 80 arcsec
Binocular amplitude of accommodation ≥ 8 D
Binocular accommodative facility with ± 2.00D lenses ≥ 5 cpm
Reading acuity ≤ 0.2 logRAD at the Radner Reading Charts
Ability to read, comprehend, and sign the informed consent form

2.2 Development of the Short Italian WRRT (SI-WRRT)

The SI-WRRT builds upon the standard Italian WRRT (Zeri et al., 2023). The latter is made up of 15 10-line passages, each line of which contains 15 high-frequency words [i.e., belonging to the 2,000 most frequent words of the Italian language; cf. "fundamental words" in De Mauro (2016)], the same words in each line, arranged in random order. The words are: *di* [of], *ha* [has], *si* [third person reflexive pronoun, used in reflexive verbs], *la* ["the" for feminine nouns], *amo* [I love], *che* [that/which], *con* [with], *era* [was], *fai* [(you) do], *non* [not], *per* [for], *una* ["a" for feminine nouns], *anno* [year], *sono* [am/are], and *uomo* [man]. Each word appears only once per line and once in a specific serial position (i.e., from 1 to 15). Additionally, the last word in each line is different from the first word in the next line, and all lines across passages are different. The typesetting conforms to the typographic specifications of the original WRRT (Wilkins et al., 1996), featuring Times New Roman font, 9-point print size (i.e., 0.5 logMAR at a viewing distance of 40 cm, whereby logMAR is the logarithm of the minimum angle of

resolution), single-spaced lines (3.15 mm), and 4-point horizontal spacing between words. For the creation of the 5-line passages, each 10-line passage from the standard Italian WRRT was split into two halves, resulting in a total of thirty 5-line passages, 15 of which (labelled with consecutive lowercase letters from "a" to "o") were used in the present study. The final layout of each passage is a paragraph 72.5 mm wide and 17.0 mm high, containing 15 words per line x 5 lines. Each passage is arranged on a separate page of a Microsoft Word file (www.microsoft.com) and printed at 1,200 dpi resolution. The set of ready-to-print passages is available in the Supplementary material.

2.3 Vision assessment

Participants underwent a preliminary comprehensive eye examination and vision assessment at the Research Centre in Optics and Optometry of the University of Milano Bicocca (COMiB). Ocular pathologies, subjective refraction, best-corrected visual acuity (BCVA), ocular motility, accommodation amplitude and facility, near point of convergence, and stereoacuity were assessed using specific standard optometric tests (whose details are reported in the Supplementary material).

Maximum reading speed (MRS), critical print size (CPS), and reading acuity (RA)—i.e., parameters that quantify near functional vision during reading (Calabrèse et al., 2016; Radner, 2016)—were measured binocularly using the standardised Italian version of the Radner Reading Charts (Radner et al., 1998; Calossi et al., 2014) at 40 cm. This test is a "sentence optotypes" chart consisting of 15 different 3-line meaningful sentences printed on cards with progressively smaller print sizes. The print size decreases logarithmically by 0.1 logMAR from the first to the 15th sentence, ranging from 1.2 to −0.2 logMAR. The number, length in characters, and frequency of use of the words are comparable across sentences, as well as syntactical construction (Radner et al., 1998). MRS is the fastest speed achieved across large print sizes representing the plateau of the reading speed curve plotted against print size, before the speed declines beyond the CPS. The CPS is the smallest print size at which one can still read at their maximum speed. RA corresponds to the smallest print size at which one can read a whole sentence. It is measured as logRAD, i.e., the logarithm of the Reading Acuity Determination (RAD), which is equivalent to the print size measured in logMAR adjusted for the reading errors made in the last sentence read entirely. Based on the outcomes of the standard optometric tests (see Section 3.1 in the Results), 29 participants kept their habitual spectacles or contact lenses, while 11 did not need any refraction correction to read the Radner Reading Charts.

2.4 SI-WRRT administration

A test and a retest session took place 2 weeks apart in the same room. Participants were assessed individually. Tests and retests were carried out using the same procedure, following detailed written instructions that were read to participants (see the file "De Luca et al. Front. Psychol. 2024 Short Italian WRRT text passages.pdf" in the Supplementary material). The same examiner, who was different from the one who carried out the vision assessment and blind to its outcome, carried out the test and retest sessions for each participant. Each passage was displayed on a single page on a reading desk at a viewing distance of 40 cm. Participants read the passages under photopic conditions (550 ± 50 lux, measured by a luxmeter HT307, HT Italia; Faenza, Italy) with an average luminance of the paper surface (eight measurements) of 135 ± 11 cd/m^2 (Chroma metre CS 100 A; Minolta; Osaka, Japan). The refraction correction for participants was the same as in the Radner Reading Charts (see section above).

Participants were asked to read the entire passage aloud as fluently and accurately as possible, with an interval of 1 min between passages. The presentation order of the passages from "a" to "o" was randomised across participants, and the reading was recorded digitally. Reading speed for correctly read words (wpm) and accuracy (percentage of reading errors) were measured offline, using Audacity (www.audacityteam.org) to replay the recordings and detect the reading onset and offset by examining the acoustic spectrum. Reading errors were scored according to the same criteria used for the 10-line passages: word substitution (replacing a word with another), word omission (skipping a word), line omission, word insertion (repeating the previously uttered word or inserting another word), and production of a non-word (a pronounceable string of letters that is not in the lexicon), with each error scored as "1."

2.5 Data analysis

The jamovi package (www.jamovi.org) was used to compute the descriptive statistics for the results of the vision assessment and the SI-WRRT (reading speed as wpm, and reading accuracy as percentage of reading errors), as well as all other analyses. The normal distribution of reading speed and accuracy data was checked using the Shapiro-Wilk test. Repeated-measures analyses were run using an ANOVA or a Friedman test, depending on the normality of the data distribution, to assess the equivalence across the 15 passages (from "a" to "o"), and any practice and fatigue effects of consecutive readings (i.e., reading order). *T*-tests were run as (two-tailed) *post-hoc* tests in case of parametric analyses. Durbin-Conover tests were run as *post-hoc* tests in case of non-parametric analyses. *Post-hoc* tests for practice and fatigue effects were run within two separate time-windows, i.e., the first seven readings and the last seven readings, respectively, based on where such effects were expected (e.g., cf. Zeri et al., 2023). For reading speed, the test-retest reliability was computed using the Intraclass Correlation Coefficient (ICC) based on a "two-way mixed effects, consistency type, single measure" model (Koo and Li, 2016), and 95% confidence intervals were calculated. Paired *t*-tests and Wilcoxon tests (for speed and accuracy data, respectively) were carried out for test-retest comparisons. Bonferroni correction for multiple testing (as calculated by jamovi) was applied in all analyses, and corrected *p*-values were reported.

3 Results

3.1 Vision assessment

All participants showed adequate eye and visual function including visual acuity, accommodation, and binocular vision,

TABLE 2 Radner Reading Charts used for assessing near functional vision during reading.

Radner Reading Charts parameter	Mean	Median	SD	Min	Max
RA (logRAD)	−0.1	−0.1	0.1	−0.2	0.1
CPS (logMAR)	0.1	0.1	0.1	−0.1	0.3
MRS (wpm)	208.4	210.1	25.7	164.3	255.9

Results indicate that participants were capable of fluently reading at near, which ensured a reliable assessment of equivalence across the SI-WRRT 15 passages.
RA, reading acuity; logRAD, logarithm of the Reading Acuity Determination (RAD), i.e., reading acuity equivalent of logMAR; CPS, critical print size; MRS, maximum reading speed; wpm, words per minute; SD, standard deviation.

as well as good near functional vision during reading. Twenty-nine participants showed negligible differences with respect to the subjective refraction measured during the optometric assessment, therefore, during both the Radner Reading Charts administration and the SI-WRRT reading session, they kept their habitual refractive correction (spectacles or contact lenses), as normally used for reading and studying. Among the remaining 11 participants, who did not wear any refractive correction, nine were emmetropes, and two had negligible myopic refractive errors that did not necessitate correction. Group results and additional details are provided in Supplementary Table 1.

Regarding the Radner Reading Charts (see Table 2), both the participants' RA (-0.1 logRAD, on average) and CPS (0.1 logMAR, on average) corresponded to a print size smaller than that of the WRRT (0.5 logMAR). Taken together, these results ensure that all participants could successfully read the SI-WRRT at their maximum speed without any limitations due to print size.

3.2 SI-WRRT

3.2.1 Equivalence across passages

Figure 1 and Table 3 present the descriptive statistics of reading speed and accuracy for the 15 passages (test session).

Reading speed showed a normal distribution for all passages (Shapiro-Wilk test: $p > 0.05$) except for passage "n" (Shapiro-Wilk test: $p = 0.005$). Reading speed across passages was 175.9 ± 3.4 wpm (range 171.4 – 183.8). Repeated-measures ANOVA revealed a statistically significant difference in reading speed across passages [$F_{(1,14)} = 3.74$; $p < 0.001$]. *Post-hoc* testing identified five significant comparisons (all p-values < 0.05 after applying Bonferroni correction). All significant comparisons involved passage "e" (paired with "h," "i," "k," "m," and "o"). In fact, passage "e" was on average 11.3 wpm (range 9.8 – 12.4 wpm) faster than these passages, a difference that is also clinically relevant (cf. Section 2.1). The average of non-significant differences with passage "e" was 6.9 wpm (range 4.8 – 10.5 wpm, see below). Despite its faster reading speed, passage "e" did not compromise accuracy, as the error rate was only 1.6% (i.e., lower than other passages). Passage "e" also showed a clinically relevant difference of 10.5 wpm with passage "l," but the comparison was not significant. Excluding passage "e" returned an average reading speed across passages of 175.3 ± 2.7 wpm (range 171.4 – 179.0).

For reading accuracy, no passage showed a normal distribution (Shapiro-Wilk test: all p-values < 0.05). The average percentage of reading errors was 2.4% ± 0.5 (median 2.5%; range 1.6 – 3.3%). The Friedman test indicated a significant difference across passages ($r = 27.4$; $p = 0.017$). *Post-hoc* testing did not show any significant paired comparison. Excluding passage "e," which showed non-equivalence in reading speed, resulted in an error rate of 2.5% ± 0.5, that is, an overall accuracy of 97.5%. Finally, reading speed was not associated with reading accuracy (Spearman's Rho = −0.14, $p = 0.387$).

3.2.2 Practice and fatigue effects

Figure 2 and Table 4 show reading speed and accuracy as a function of reading order (test session).

Reading speed showed a normal distribution for all readings (Shapiro-Wilk test: $p > 0.05$) except for the 1st, 4th, 5th, and 7th readings (Shapiro-Wilk test: $p < 0.05$). Reading speed across all readings was 175.9 ± 3.02 wpm (range 169.5 – 179.6). Repeated-measures ANOVA revealed a significant difference across readings [$F_{(1,14)} = 2.94$; $p < 0.001$], indicating an effect of reading order. As regards the practice effect, *post-hoc* testing in the first time-window (i.e., across the first seven readings) identified a single significant comparison between the 1st and the 5th reading (whereby the 1st reading was slower than the 5th; $p = 0.047$ after applying Bonferroni correction), although the difference (8.2 wpm) was not clinically relevant. As regards the fatigue effect, *post-hoc* testing in the second time-window (i.e., across the last seven readings) identified two significant comparisons: between the 13th and the 10th reading, and between the 13th and the 15th reading, whereby the 13th was slower than both (see Table 4; $p < 0.001$ and $p = 0.021$, respectively after applying Bonferroni correction). However, only the difference between the 13th and the 10th reading was also clinically relevant. Reading speed between the 3rd and the 10th reading (i.e., the plateau visible in Figure 2A, before performance gets slower) was 177.9 ± 1.9 wpm (range 174.8 – 179.6).

As regards reading accuracy, none of the readings showed a normal distribution (Shapiro-Wilk test: all p-values < 0.01). The percentage of reading errors was 2.4% ± 0.6 (median 2.3%; range 1.5 – 3.8%). Friedman test revealed a significant difference across readings ($r = 30.9$; $p = 0.006$), but *post-hoc* testing showed no significant comparisons in either time-windows.

3.2.3 Test-retest reliability

All participants underwent a retest 2 weeks after the initial session. Figure 3 shows the results of both test and retest sessions, presenting data about passages from "a" to "o," separately for reading speed and accuracy. Descriptive statistics and paired comparisons between test and retest sessions for reading speed and accuracy, along with Intraclass Correlation Coefficients (ICC) specifically for reading speed, are provided in Table 5. Generally, retest performance showed an improvement in both reading speed (180.7 ± 3.6 wpm) and accuracy (1.6% ± 0.3). Reading speed was faster in 14 out of the 15 passages, with a significant increase observed only in one passage ("l;" $p = 0.014$ after applying Bonferroni correction), which also represented a clinically relevant difference (10.3 wpm). The other differences were neither

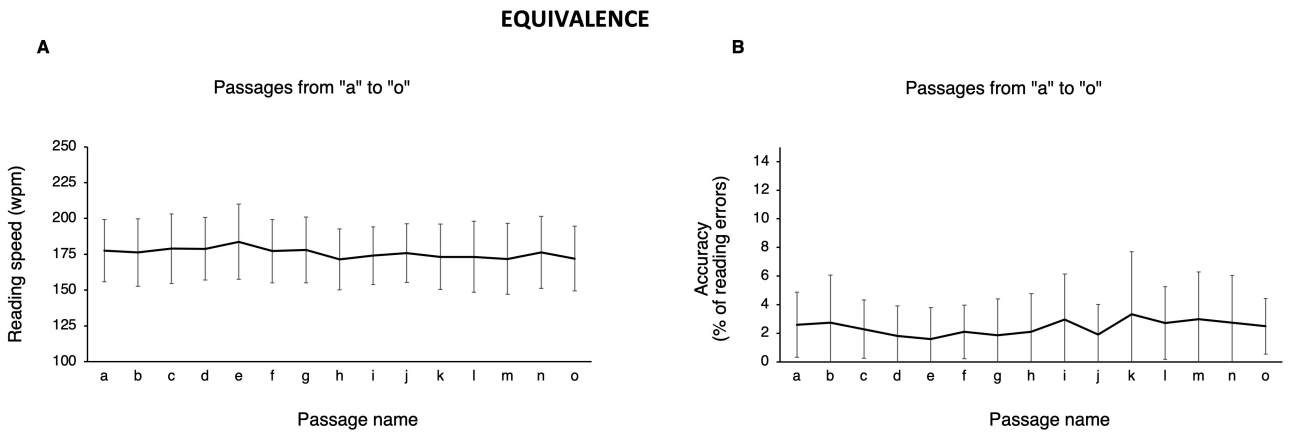


FIGURE 1
Equivalence across passages. Mean (error bars: standard deviation) of **(A)** reading speed (words per minute; wpm) and **(B)** accuracy (% of reading errors) for the 15 passages ("a"—"o").

TABLE 3 Equivalence across passages.

	Reading speed (wpm)				Reading accuracy (% of reading errors)				
Passage name	Mean	SD	Min	Max	Mean	Median	SD	Min	Max
a	177.5	21.7	135.1	220.4	2.6	2.4	2.3	0.0	10.9
b	176.2	23.7	134.6	223.1	2.7	1.7	3.3	0.0	15.2
c	179.0	24.2	129.2	231.9	2.3	1.7	2.0	0.0	10.3
d	178.9	21.8	136.1	232.6	1.8	1.4	2.1	0.0	10.9
e	183.8	26.3	138.3	239.9	1.6	1.0	2.2	0.0	9.7
f	177.3	22.1	129.9	236.5	2.1	2.0	1.9	0.0	6.9
g	178.0	23.0	129.0	242.1	1.9	1.3	2.5	0.0	8.6
h	171.4	21.2	131.2	226.1	2.1	1.4	2.7	0.0	14.6
i	174.0	20.1	137.7	223.1	2.9	2.1	3.2	0.0	10.9
j	175.8	20.6	134.5	224.6	1.9	1.4	2.1	0.0	8.0
k	173.2	22.8	125.2	233.3	3.3	1.7	4.4	0.0	25.0
l	173.3	24.7	117.0	229.9	2.7	2.0	2.5	0.0	9.4
m	171.8	24.8	120.0	228.0	3.0	2.1	3.3	0.0	16.2
n	176.4	25.2	126.4	251.8	2.7	2.0	3.3	0.0	16.2
o	172.0	22.5	128.1	225.4	2.5	2.7	1.9	0.0	7.0

Mean, standard deviation (SD), and range (Min – Max) for reading speed (words per minute; wpm) and accuracy (% of reading errors) for the 15 passages ("a"—"o").

statistically significant, nor clinically relevant. Accuracy improved for all passages, but only three ("a," "l," and "o") showed a significant difference (1.2%, $p = 0.008$; 1.6%, $p = 0.004$; 1.1%, $p = 0.041$, respectively). The ICC for reading speed (wpm), the standard metric for WRRT, indicated moderate-to-good reliability for all passages (range 0.67 – 0.82).

4 Discussion

Reading performance on parallel forms of texts is commonly used by clinicians and researchers in vision science as a tool to reliably assess (by measuring reading speed) the effectiveness of interventions in patients with vision deficits, developmental reading disabilities, or visual perception impairments following acquired brain lesion (e.g., Bailey and Lakshminaryanan, 1997; Spitzyna et al., 2007; Tilanus et al., 2019). Recently, it has been proposed that the WRRT may be considered as a RAN (rapid automatized naming) test (Gilchrist et al., 2021), i.e., a task commonly used in neuropsychology of developmental reading disorders. Both tests share rapid processing from left to right of arrays of recurrent familiar stimuli, although WRRT is based on reading recurrent unrelated words, while RAN is based on naming recurrent items (digits, coloured squares, or other visual

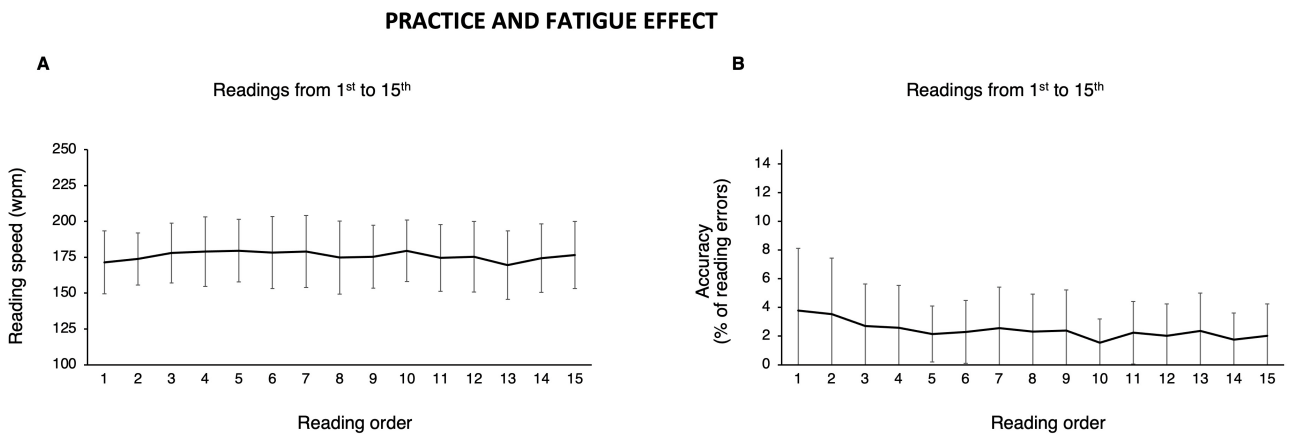


FIGURE 2
Practice and fatigue effects. Mean (error bars: standard deviation) of **(A)** reading speed (words per minute; wpm) and **(B)** accuracy (% of reading errors) as a function of the reading order.

TABLE 4 Practice and fatigue effects.

	Reading speed (wpm)				Reading accuracy (% of reading errors)				
Reading order	Mean	SD	Min	Max	Mean	Median	SD	Min	Max
1st reading	171.4	21.9	138.3	231.0	3.8	2.7	4.3	0.0	25.0
2nd reading	173.7	18.2	139.2	218.9	3.5	2.1	3.9	0.0	16.2
3rd reading	177.9	20.7	129.2	228.0	2.7	2.1	2.9	0.0	16.2
4th reading	178.9	24.2	138.1	232.6	2.6	1.4	3.0	0.0	13.4
5th reading	179.6	21.9	146.0	239.9	2.1	1.7	2.0	0.0	8.0
6th reading	178.3	25.0	134.5	236.5	2.3	2.0	2.2	0.0	7.0
7th reading	179.1	25.2	137.7	242.1	2.5	2.0	2.9	0.0	10.9
8th reading	174.8	25.5	134.0	231.9	2.3	1.4	2.6	0.0	11.8
9th reading	175.3	21.9	129.0	227.5	2.4	1.4	2.8	0.0	10.9
10th reading	179.5	21.5	129.9	235.5	1.5	1.3	1.6	0.0	4.9
11th reading	174.5	23.3	128.1	223.1	2.2	1.7	2.2	0.0	9.3
12th reading	175.4	24.6	126.4	251.8	2.0	1.4	2.2	0.0	8.6
13th reading	169.5	23.9	120.0	222.0	2.4	2.1	2.6	0.0	14.6
14th reading	174.4	24.0	117.0	229.9	1.8	1.3	1.9	0.0	7.0
15th reading	176.6	23.4	125.2	233.3	2.0	1.3	2.2	0.0	7.1

Mean, standard deviation (SD), and range (Min – Max) for reading speed (words per minute; wpm) and accuracy (% of reading errors) for the readings, from the 1st to the 15th.

stimuli arranged in arrays; Denckla and Rudel, 1974). The proposal has a heuristic value, since studies in the field of developmental neuropsychology reported that RAN tasks are associated with reading fluency (e.g., Georgiou et al., 2013; Landerl et al., 2019) and reading deficits (e.g., Denckla and Rudel, 1976; Norton and Wolf, 2012). In ophthalmology and optometry, reading performance allows the evaluation of reading acuity, critical print size, reading speed, and maximum reading speed. These parameters are used to measure the outcomes of interventions such as cataract surgery with lens implantation, presbyopia correction, determination of magnification need under different visual conditions and low vision aids, prismatic corrections, eye exercises, or refractive modifications (Alió et al., 2011; Buckhurst et al., 2012; Crossland et al., 2019; O'Leary and Evans, 2006; Zeri et al., 2018). These parameters are also commonly adopted to evaluate interventions for the improvement of reading performance in patients with developmental dyslexia (Tilanus et al., 2019; Wilkins et al., 1996) and hemianopia (Daibert-Nido et al., 2021; see also Schuett et al., 2008 for a review on hemianopic dyslexia).

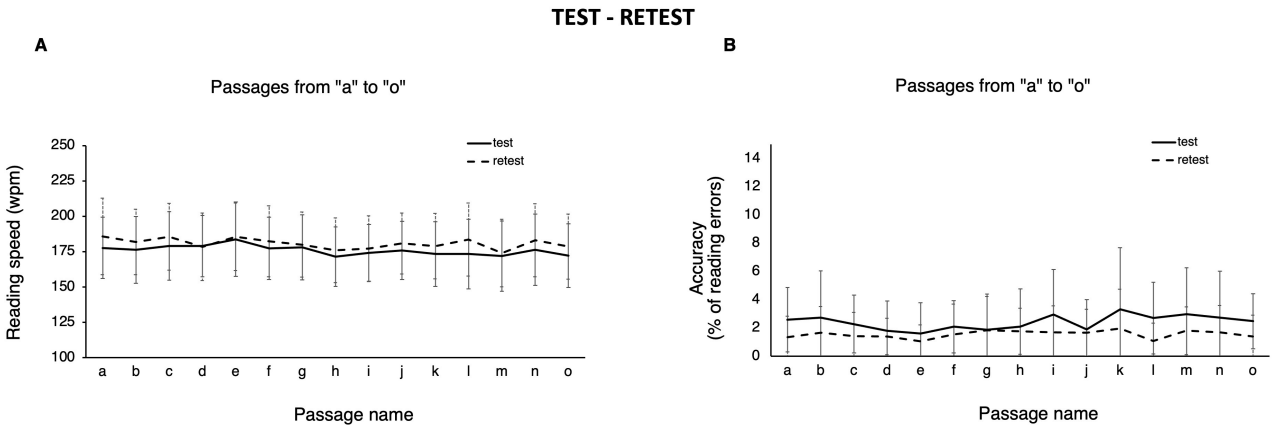


FIGURE 3
Test-retest. Mean (error bars: standard deviation) of **(A)** reading speed (words per minute; wpm) and **(B)** accuracy (% of reading errors) for the 15 passages ("a"—"o"), separately for test (continuous line) and retest (dashed line) sessions.

TABLE 5 Test-retest reliability separately for reading speed (words per minute; wpm) and accuracy (% of reading errors).

	Reading speed (wpm)				Reading accuracy (% of reading errors)		
Passage name	Test Mean ± SD	Retest Mean ± SD	Paired *t*-test comparison	ICC (95% CI)	Test Mean ± SD	Retest Mean ± SD	Wilcoxon test comparison
a	177.5 ± 21.7	185.8 ± 27.1	n.s.	0.74* (0.56 – 0.85)	2.6 ± 2.3	1.4 ± 1.5	***p* = 0.008**
b	176.2 ± 23.7	181.8 ± 23.2	n.s.	0.78* (0.62 – 0.88)	2.7 ± 3.3	1.7 ± 1.8	n.s.
c	179.0 ± 24.2	185.5 ± 23.7	n.s.	0.73* (0.55 – 0.85)	2.3 ± 2.0	1.4 ± 1.7	n.s.
d	178.9 ± 21.8	178.3 ± 24.0	n.s.	0.76* (0.60 – 0.87)	1.8 ± 2.1	1.4 ± 1.3	n.s.
e	183.8 ± 26.3	185.5 ± 23.9	n.s.	0.68* (0.47 – 0.82)	1.6 ± 2.2	1.1 ± 1.2	n.s.
f	177.3 ± 22.1	182.3 ± 25.2	n.s.	0.74* (0.56 – 0.86)	2.1 ± 1.9	1.6 ± 2.1	n.s.
g	178.0 ± 23.0	179.9 ± 23.0	n.s.	0.82* (0.68 – 0.90)	1.9 ± 2.5	1.8 ± 2.4	n.s.
h	171.4 ± 21.2	176.0 ± 23.0	n.s.	0.81* (0.67 – 0.89)	2.1 ± 2.7	1.8 ± 1.6	n.s.
i	174.0 ± 20.1	177.1 ± 23.3	n.s.	0.81* (0.67 – 0.90)	2.9 ± 3.2	1.7 ± 1.9	n.s.
j	175.8 ± 20.6	180.7 ± 21.6	n.s.	0.69* (0.48 – 0.82)	1.9 ± 2.1	1.7 ± 1.6	n.s.
k	173.2 ± 22.8	178.8 ± 23.1	n.s.	0.73* (0.54 – 0.85)	3.3 ± 4.4	2.0 ± 2.8	n.s.
l	173.3 ± 24.7	183.6 ± 25.8	***p* = 0.014**	0.74* (0.56 – 0.86)	2.7 ± 2.5	1.1 ± 1.3	***p* = 0.004**
m	171.8 ± 24.8	173.9 ± 23.8	n.s.	0.72* (0.53 – 0.84)	3.0 ± 3.3	1.8 ± 1.7	n.s.
n	176.4 ± 25.2	183.0 ± 25.9	n.s.	0.67* (0.45 – 0.81)	2.7 ± 3.3	1.7 ± 1.9	n.s.
o	172.0 ± 22.5	178.5 ± 23.0	n.s.	0.77* (0.61 – 0.87)	2.5 ± 1.9	1.4 ± 1.5	***p* = 0.041**

The table reports descriptive statistics (mean and standard deviations) and p-values of paired comparisons between test and retest. Intraclass Correlation Coefficient (ICC) and confidence intervals (CIs) between test and retest (calculated with two-way mixed effects model, consistency, and single measure) are presented for reading speed—the standard metric for WRRT. Significant p-values correspond to corrected p-values after applying Bonferroni correction.
*ICC significant with $p < 0.001$.

4.1 Parallel forms of texts in vision science

The most common tests measuring these parameters are serial texts such as the Bailey-Lovie Reading Sentence Chart (Bailey and Lovie, 1980), MNRead Acuity Charts (Mansfield et al., 1993; Ahn et al., 1995), and Radner Reading Charts (Radner et al., 1998). These tests are based on very short texts (1–3 lines) made up of either unrelated words (as in the Bailey-Lovie chart) or continuous text (i.e., sentences with a meaning, as in the MNRead and Radner charts), printed with progressively smaller print size. Other common materials suitable for repeated measurements of reading speed are represented by longer texts, such as the passages of the New International Reading Speed Texts (IreST; Hahn et al., 2006; Trauzettel-Klosinski et al., 2012), which are printed with a fixed print size, and another version of the Radner charts made up of long paragraphs (i.e., texts longer than the sentences of the original charts; Radner et al., 2016). Standardised versions of Radner, MNRead, and IreST exist in different languages (see

Rubin, 2013 and Radner, 2017, for reviews). Additionally, the IreST is matched for psycholinguistic variables and syntactic complexity across languages.

All tests have advantages and disadvantages. Short texts with a progressive reduction in print size are commonly used to determine the outcome of treatments and interventions, as mentioned above. They accurately assess CPS and RA very quickly, but may be less accurate in measuring speed, unless performance is digitally recorded, or examiners are thoroughly trained (see Radner et al., 2017 for a discussion on the accuracy of reading time measurements). It has been observed that short texts (e.g., texts no longer than 60 characters; cf. Rubin, 2013) may inaccurately measure reading speed due to several factors, including an examiner's reaction time in starting and stopping the stopwatch, as well as pauses and self-corrections by the reader. Long texts are commonly adopted to measure sustained reading and functional reading speed (typically assessed in low vision patients). It has been claimed that long texts yield more reliable reading speed measures and should therefore be preferred in repeated measurements (e.g., Kortuem et al., 2021). However, even after thorough linguistic matching of text complexity and equivalence in the number of characters and syllables, number and length of words, as well as words position and overall layout, long texts may still not yield comparable reading measures. In fact, differences may remain undetected unless the texts are statistically validated (e.g., Brussee et al., 2015). For example, Radner et al. (2016) found unexpected results (i.e., non-comparable reading speeds) in the development of texts for their long paragraphs that were built to have equivalent readability scores. It is possible that other cognitive factors (incl. emotional and attentional factors; cf. Radner et al., 2016) played a role simply because of the presence of syntax, semantics, and text meaning. In other words, even if most texts are very simple and suitable for 6th-grade readers (e.g., Trauzettel-Klosinski et al., 2012), the minimal literacy demand may not be sufficient to avoid uncontrolled effects, because the presence of a semantic context is potentially capable of influencing reading speed (cf. Rubin, 2013, about the semantic context controversy). In addition, it has been shown that short equivalent MNRead sentences generated by algorithms under strict linguistic and layout constraints may determine non-comparable reading performances, thus prompting a recommendation to screen new sentences before using them (Mansfield et al., 2019).

Therefore, matching linguistic variables makes it challenging to generate and validate a series of parallel meaningful passages. An interesting solution adopted in vision science is to neutralise/minimise the role of syntax and semantics at the sentence-level by using the same set of shuffled unrelated words across lines and passages. This approach was introduced in the original WRRT by Wilkins et al. (1996), where 15 high frequency words are randomised across 10-line 15-word passages. In this vein, the standard Italian WRRT (Zeri et al., 2023) was generated by transliterating (not directly translating) the original WRRT and expanded the number of passages, demonstrating the equivalence of 15 10-line passages of unrelated words. Therefore, the standard Italian WRRT provided suitable material for studies with repeated-measures designs involving multiple conditions or measurements over time (e.g., baseline and follow-ups). The reading speed observed in Italian participants who read the standard Italian WRRT (167.3 $\pm$ 1.6 wpm) was consistent with the results obtained in studies measuring reading speed (see Brysbaert, 2019, for a meta-analysis based on data from several languages; see also Gilchrist et al., 2021). The standard Italian WRRT showed a practice effect, which expired after the first reading, while there was no fatigue effect. However, studies involving the elderly or brain-damaged patients may be vulnerable to fatigue when using long texts to test multiple conditions. For instance, each passage of the standard Italian WRRT (unrelated words) takes about 55 s to be read, while each passage of the Italian version of the IReST (meaningful sentences) takes about 45 s. Therefore, the need for a shorter WRRT for studies involving patients who may get tired more rapidly and/or have attentional deficits has prompted the development of a shorter test, the SI-WRRT (the focus of the present study), based on the same principles and constraints adopted in the original WRRT (as well as in the standard Italian WRRT).

4.2 Equivalence across SI-WRRT passages

The present study showed that the passages of the SI-WRRT are equivalent to each other, except one. Specifically, passage "e" showed a faster reading speed compared to the others (see Table 3). Such discrepancy was both statistically significant and clinically relevant with respect to five passages, and only clinically relevant with respect to another one. However, the average reading speed across passages remains comparable, whether including passage "e" (175.9 wpm) or excluding it (175.3 wpm) from the testing set. Furthermore, there is no speed/accuracy trade-off, as passage "e" was not read faster at the expense of accuracy. Overall, there is no association between reading speed and accuracy in the whole SI-WRRT. Indeed, accuracy was higher (but not statistically different) for passage "e" (1.6% error rate). Therefore, as passage "e" is the only non-equivalent one, we strongly recommend excluding it from the items used for repeated measures, and instead using it as a familiarisation item (see practice effect in Section 4.3). This finding challenges the notion that simply rearranging the positions of unrelated words automatically generates equivalent passages (Wilkins et al., 1996; Gilchrist et al., 2021). It also underscores the importance of testing the equivalence of reading speed and accuracy for newly generated passages of unrelated words, even if they are excerpts derived from previously tested and validated passages (as it is the case with the SI–WRRT, which was derived from the standard Italian WRRT), before using them in research or in clinical practice.

As mentioned above, the average reading speed observed for a passage of the SI-WRRT is 175.3 wpm. This value corresponds to a reading time of 25.5 s, which is about half the time needed for the 10-line version (54.9 s). A session of 14 passages is completed in ~19 min (including 1 min of rest between passages). The difference in reading speed between the 5-line and the 10-line test is 8.6 wpm, a value below the clinically relevant difference (i.e., 10 wpm). Indeed, both the 5-line and 10-line speed values lie within the range indicated in the above-mentioned review by Brysbaert (2019). However, the slightly faster speed for the 5-line passages

may be explained by the reduction in visual crowding in the test. The original idea of Wilkins et al. (1996) was to create a test that maximised visual stress (with crowded word arrangement on a line, and a tight line spacing), along with neutralising/minimising syntactical and semantical implications. Reducing the number of lines in the test from 10 to 5 could have diminished visual stress by reducing the layout size of the text, and thus the density of the page. This hypothesis could be tested in future studies by adding flankers (i.e., lines of text that enhance the density of the layout, but are not read by participants).

Differently from the vision assessment in the 10-line study, in the present study we also administered the Italian version of the Radner Reading Charts (Calossi et al., 2014) to quantify near functional vision during reading. The participants' RA and CPS were better than the visual capacity required by the WRRT print-size, which is supra-threshold (0.5 logMAR) with respect to the reading acuity (-0.1 logRAD) and CPS (0.1 logMAR) assessed in our sample. Therefore, using this test allowed us to ascertain that participants were in optimal visual reading conditions and could read the WRRT at their maximum speed without any limitations due to the print size of passages.

The accuracy is almost identical in the two versions (2.5% ± 0.3 and 2.4% ± 0.5 error rates, for standard WRRT and SI-WRRT, respectively, corresponding to <4 and <2 words, respectively). Reading accuracy measurements should always be part of a protocol assessing reading speed. While reading errors are accounted for by default when reading optotypes (as the criterion to go from a given print size to a smaller one relies on correctly reading a sentence), long readings need to score reading errors during online performance. However, this has not always been accomplished in the past (see Brussee et al., 2014, for a review). In the present study, reading speed was determined by analysing digital audio recordings, which enabled an accurate measurement of both reading errors and reading time. This allowed a reliable computation of reading speed based on correctly read words, as per WRRT principles (Wilkins et al., 1996). In the present study, error rate was very low (range 1.6 – 3.3%). Since error rate may not be negligible in patients with low vision, neuropsychological deficits due to acquired brain lesions, or developmental reading disorders, measuring reading errors should always be part of the procedure to reliably measure reading speed.

4.3 Practice and fatigue effects

Contrary to the findings reported in the standard Italian WRRT (Zeri et al., 2023), this study revealed a less pronounced practice effect for reading speed. The only significant difference occurred between the 1st and the 5th reading, which was not clinically relevant. Overall, results indicate a slow reading speed improvement characterised by progressively faster speeds across the initial five readings, with the first two readings slower than the following ones, and a plateau from the 3rd reading onwards (see Figure 2A and Table 4). One may presume that in vulnerable populations, such as the elderly or neuropsychological patients, statistically significant differences may emerge more readily (see section 4.5). Hence, we confirm the need to familiarise with the test (i.e., reading at least one passage) before proceeding with consecutive readings for experimental or clinical purposes (i.e., repeated measures). Therefore, passage "e" can be used to this purpose (see Section 4.2). This prevents from biassing the interpretation of the effect of interventions (e.g., Allen et al., 2012). As regards accuracy, there were neither significant effects, nor clinically relevant differences.

As for the fatigue effect, significant differences occurred between the 10th and the 13th reading, and between the 13th and the 15th (only the former being clinically relevant). Figure 2A shows a progressive decline in reading speed (mostly non-significant) evident after the 10th reading (179.5 wpm) down to the dip by the 13th reading (169.5 wpm). The last two readings of Figure 2A can be interpreted as a kind of "relief-effect" [an identical trend, although not significant, was found in Zeri et al. (2023)]. Therefore, both the 10- and the 5-line versions are potentially suitable to assess prolonged reading for at least 10 consecutive readings in the same session. Once more, studies on the elderly or patients with specific issues are needed to determine fatigue effects in populations different from the one tested in the present study (i.e., healthy controls; see Section 4.5).

4.4 Test-retest reliability

In line with previous research on the WRRT (e.g., Stifter et al., 2004), the test-retest reliability was assessed. In both the present study and in previous ones (e.g., Wilkins et al., 1996; Zeri et al., 2023), results showed a generally improved performance in both reading speed and accuracy at retest. This can be easily explained by a slight learning process. Importantly, as regards reading speed, which is the standard metric for WRRT, 14 passages were neither significantly faster, nor showed a clinically relevant difference (i.e., >10 wpm) between test and retest. One passage ("l") was significantly faster at retest, and the difference (10.3 wpm) was clinically relevant. As regards accuracy, only three passages ("a," "l," and "o") improved significantly, but the differences were negligible (range 1.1–1.6% of errors). Test-retest reliability showed moderate-to-good ICC for all passages, with the exception of passages "e," "j," and "n" (ICC = 0.68, 0.69, and 0.67, respectively). However, we would like to point out that, although correlations are usually computed in studies assessing the reliability of parallel forms across time, we suspect that such analysis may not be fully appropriate to assess the validity of texts for repeated measures, as in the present context (see Radner et al., 2016, for a similar position). This is because the availability of many equivalent texts makes it highly unlikely that an examiner would need to resort to the very same item twice at any time point. In other words, a passage would not be presented twice, since other equivalent passages are available. Therefore, as consistency across time depends on equivalence, assessing test-retest reliability would probably have little practical relevance.

4.5 Limitations and future directions

A possible limitation to the present study is that the trend observed in the practice and fatigue effect cannot be generalised to populations different from the sample tested here. Indeed, our

study required the selection of participants with optimal reading capacity to determine the equivalence of the passages, and hence it was conducted on young, healthy individuals. However, the absence of a pronounced fatigue effect in our sample of readers does not prevent the potential occurrence of fatigue phenomena in more vulnerable populations (such as the elderly or neuropsychological patients, as well as readers with low vision, impaired reading ability, or attentional deficits). Therefore, it is essential that future studies will test the present materials in such populations.

Furthermore, future studies may test the hypothesis that the layout of the SI-WRRT leads to a reduction in reading time compared to the standard Italian WRRT because it decreases the density of the page, thereby reducing crowding. The reading materials to test this hypothesis may be based on the addition of flankers above and below the main text (i.e., lines of text that enhance the density of the layout, but are not read by participants).

5 Conclusions

The present study confirms the equivalence of 14 passages in the SI-WRRT, highlighting its usefulness for assessing reading speed in the elderly or neuropsychological patients in repeated-measures designs, due to the halved reading times of the 5-line passages with respect to the 10-line ones. As already suggested in Zeri et al. (2023), we reiterate the recommendation of providing a familiarisation item (i.e., giving participants a first item, which is not part of the assessment) before proceeding with the actual test for experimental or clinical purposes. Importantly, the non-equivalence of one passage underscores the need of a formal statistical validation before adopting random rearrangements of words to generate new passages.

Data availability statement

The datasets presented in this article are not readily available because of restrictions specified in the study consent-form, and conditions for approval from the local Ethics Committee, concerning participant confidentiality and privacy. Requests to access the datasets should be directed to the corresponding author, giuliacarlotta.rizzo@unimib.it.

Ethics statement

The studies involving humans were approved by Ethics Committee of the University of Milano-Bicocca. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MDL: Data curation, Formal analysis, Methodology, Supervision, Validation, Visualisation, Writing – original draft, Writing – review & editing. DN: Methodology, Validation, Writing – original draft, Writing – review & editing. GCR: Data curation, Investigation, Writing – review & editing. RD: Resources, Writing – review & editing. ST: Resources, Supervision, Writing – review & editing. FZ: Conceptualisation, Data curation, Formal analysis, Methodology, Project administration, Resources, Supervision, Validation, Visualisation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was partially supported by the Italian Ministry of Health (Ministero della Salute, Ricerca Corrente, Linea 1) to IRCCS Fondazione Santa Lucia.

Acknowledgments

The authors wish to thank Tomas Perego for helping in data collection.

Conflict of interest

The 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1448817/full#supplementary-material

References

Ahn, S. J., Legge, G. E., and Luebker, A. (1995). Printed cards for measuring low-vision reading speed. *Vis. Res.* 35, 1939–1944. doi: 10.1016/0042-6989(94)00294-V

Alió, J. L., Plaza-Puche, A. B., Piñero, D. P., Amparo, F., Jiménez, R., Rodríguez-Prats, J. L., et al. (2011). Optical analysis, reading performance, and quality-of-life

evaluation after implantation of a diffractive multifocal intraocular lens. *J. Cataract Refract. Surg.* 37, 27–37. doi: 10.1016/j.jcrs.2010.07.035

Allen, P. M., Dedi, S., Kumar, D., Patel, T., Aloo, M., and Wilkins, A. J. (2012). Accommodation, pattern glare, and coloured overlays. *Perception* 41, 1458–1467. doi: 10.1068/p7390

Altpeter, E. K., Marx, T., Nguyen, N. X., Naumann, A., and Trauzettel-Klosinski, S. (2015). Measurement of reading speed with standardized texts: a comparison of single sentences and paragraphs. *Graefes Arch. Clin. Exp. Ophthalmol.* 253, 1369–1375. doi: 10.1007/s00417-015-3065-4

Bailey, I. L., and Lovie, J. E. (1980). The design and use of a new near-vision chart. *Am. J. Optom. Physiol. Optics* 57, 378–387. doi: 10.1097/00006324-198006000-00011

Bailey, J. E., and Lakshminaryanan, V. (1997). "Assessing reading ability in normal and low vision using the mnread reading acuity chart: preliminary results," in *Basic and Clinical Applications of Vision Science. Documenta Ophthalmologica Proceedings Series, Vol 60*, ed. V. Lakshminarayanan (Dordrecht: Springer), 247–250.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Arch. Clin. Neuropsychol.* 20, 517–529. doi: 10.1016/j.acn.2004.12.003

Brussee, T., van Nispen, R. M., Klerkx, E. M., Knol, D. L., and van Rens, G. H. (2015). Comparison of reading performance tests concerning difficulty of sentences and paragraphs and their reliability. *Ophthalmic Physiol. Opt.* 35, 324–335. doi: 10.1111/opo.12204

Brussee, T., van Nispen, R. M., and van Rens, G. H. (2014). Measurement properties of continuous text reading performance tests. *Ophthalmic Physiol. Opt.* 34, 636–657. doi: 10.1111/opo.12158

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *J. Mem. Lang.* 109:104047. doi: 10.1016/j.jml.2019.104047

Buckhurst, P. J., Wolffsohn, J. S., Gupta, N., Naroo, S. A., Davies, L. N., and Shah, S. (2012). Development of a questionnaire to assess the relative subjective benefits of presbyopia correction. *J. Cataract Refract. Surg.* 38, 74–79. doi: 10.1016/j.jcrs.2011.07.032

Calabrèse, A., Cheong, A. M., Cheung, S. H., He, Y., Kwon, M., Mansfield, J. S., et al. (2016). Baseline MNREAD measures for normally sighted subjects from childhood to old age. *Invest. Ophthalmol. Vis. Sci.* 57, 3836–3843. doi: 10.1167/iovs.16-19580

Calossi, A., Boccardo, L., Fossetti, A., and Radner, W. (2014). Design of short Italian sentences to assess near vision performance. *J. Optom.* 7, 203–209. doi: 10.1016/j.optom.2014.05.001

Crossland, M. D., Starke, S. D., Imielski, P., Wolffsohn, J. S., and Webster, A. R. (2019). Benefit of an electronic head-mounted low vision aid. *Ophthal. Physiol. Opt.* 39, 422–431. doi: 10.1111/opo.12646

Daibert-Nido, M., Pyatova, Y., Cheung, K., Nayomi, C., Markowitz, S. N., Bouffet, E., et al. (2021). Case report: visual rehabilitation in hemianopia patients. Home-based visual rehabilitation in patients with hemianopia consecutive to brain tumor treatment: feasibility and potential effectiveness. *Front. Neurol.* 12:680211. doi: 10.3389/fneur.2021.680211

De Mauro, T. (2016). *Nuovo vocabolario di base della lingua italiana*. Available at: https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana (accessed May 15, 2018).

Denckla, M. B., and Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters and numbers by normal children. *Cortex* 10, 186–202. doi: 10.1016/S0010-9452(74)80009-2

Denckla, M. B., and Rudel, R. (1976). Naming of object-drawings by dyslexic and other learning disabled children. *Brain. Lang.* 3, 1–15. doi: 10.1016/0093-934X(76)90001-8

Evans, B. J. W., and Joseph,. F. (2002). The effect of coloured filters on the rate of reading in an adult student population. *Ophthal. Physiol. Opt.* 22, 535–545. doi: 10.1046/j.1475-1313.2002.00071.x

Georgiou, G. K., Parrila, R., Cui, Y., and Papadopoulos, T. C. (2013). Why is rapid automatized naming related to reading? *J. Exp. Child. Psychol.* 115, 218–225. doi: 10.1016/j.jecp.2012.10.015

Gilchrist, J. M., Allen, P. M., Monger, L., Srinivasan, K., and Wilkins, A. (2021). Precision, reliability and application of the Wilkins Rate of Reading Test. *Ophthal. Physiol. Opt.* 41, 1198–1208. doi: 10.1111/opo.12894

Hahn, G. A., Penka, D., Gehrlich, C., Messias, A., Weismann, M., Hyvärinen, L., et al. (2006). New standardised texts for assessing reading performance in four European languages. *Br. J. Ophthalmol.* 90, 480–484. doi: 10.1136/bjo.2005.087379

Kaltenegger, K., Kuester, S., Altpeter-Ott, E., Eschweiler, G. W., Cordey, A., Ivanov, I. V., et al. (2019). Effects of home reading training on reading and quality of life in AMD-a randomized and controlled study. *Graefes Arch. Clin. Exp. Ophthalmol.* 257,1499–1512. doi: 10.1007/s00417-019-04328-9

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kortuem, C., Marx, T., Altpeter, E. K., Trauzettel-Klosinski, S., and Kuester-Gruber, S. (2021). Comparing reading speeds for reading standardized single sentences and paragraphs in patients with maculopathy. *Ophthal. Res.* 64, 512–522. doi: 10.1159/000509687

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., et al. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Sci. Stud. Read.* 23, 220–234. doi: 10.1080/10888438.2018.1510936

Ludlow, A. K., Wilkins, A. J., and Heaton, P. (2006). The effect of coloured overlays on reading ability in children with autism. *J. Autism Dev. Disord.* 36, 507–516. doi: 10.1007/s10803-006-0090-5

Mansfield, J. S., Ahn, S. J., Legge, G. E., and Luebker, A. (1993). "A new reading acuity chart for normal and low vision," in *Noninvasive Assessment of the Visual System, Technical Digest Series* (Washington D.C.: Optica Publishing Group), NSuD.3.

Mansfield, J. S., Atilgan, N., Lewis, A. M., and Legge, G. E. (2019). Extending the MNREAD sentence corpus: computer-generated sentences for measuring visual performance in reading. *Vis. Res.* 158, 11–18. doi: 10.1016/j.visres.2019.01.010

Möller, M. C., Nygren de Boussard, C., Oldenburg, C., and Bartfai, A. (2014). An investigation of attention, executive, and psychomotor aspects of cognitive fatigability. *J. Clin. Exp. Neuropsychol.* 36, 716–729. doi: 10.1080/13803395.2014.933779

Monger, L., Wilkins, A., and Allen, P. (2015). Identifying visual stress during a routine eye examination. *J. Optom.* 8,140–145. doi: 10.1016/j.optom.2014.10.001

Northway, N. (2003). Predicting the continued use of overlays in school children—a comparison of the Developmental Eye Movement test and the Rate of Reading test. *Ophthalmic Physiol. Opt.* 23, 457–464. doi: 10.1046/j.1475-1313.2003.00144.x

Norton, E. S., and Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: implications for understanding and treatment of reading disabilities. *Annu. Rev. Psychol.* 63, 427–452. doi: 10.1146/annurev-psych-120710-100431

O'Leary, C. I., and Evans, B. J. W. (2006). Double-masked randomised placebo-controlled trial of the effect of prismatic corrections on rate of reading and the relationship with symptoms. *Ophthal. Physiol. Opt.* 26, 555–565. doi: 10.1111/j.1475-1313.2006.00400.x

Radner, W. (2016). Near vision examination in in presbyopia patients: do we need good homologated near vision charts? *Eye Vis.* 3:7. doi: 10.1186/s40662-016-0061-7

Radner, W. (2017). Reading charts in ophthalmology. *Graefes Arch. Clin. Exp. Ophthalmol.* 255, 1465–1482. doi: 10.1007/s00417-017-3659-0

Radner, W., Diendorfer, G., Kainrath, B., and Kollmitzer, C. (2017). The accuracy of reading speed measurement by stopwatch versus measurement with an automated computer program (rad-rd©). *Acta Ophthalmol.* 95, 211–216. doi: 10.1111/aos.13201

Radner, W., Radner, S., and Diendorfer, G. (2016). A new principle for the standardization of long paragraphs for reading speed analysis. *Graefes Arch. Clin. Exp. Ophthalmol.* 254, 177–184. doi: 10.1007/s00417-015-3207-8

Radner, W., Willinger, U., Obermayer, W., Mudrich, C., Velikay-Parel, M., and Eisenwort, B. (1998). A new reading chart for simultaneous determination of reading vision and reading speed. Klin. Monbl. Augenheilkd. 213, 174–181. German. doi: 10.1055/s-2008-1034969

Rubin, G. S. (2013). Measuring reading performance. *Vis. Res.* 20, 43–51. doi: 10.1016/j.visres.2013.02.015

Schuett, S., Heywood, C. A., Kentridge, R. W., and Zihl, J. (2008). The significance of visual information processing in reading: insights from hemianopic dyslexia. *Neuropsychologia* 46, 2445–2462. doi: 10.1016/j.neuropsychologia.2008.04.016

Spitzyna, G. A., Wise, R. J., McDonald, S. A., Plant, G. T., Kidd, D., Crewes, H., et al. (2007). Optokinetic therapy improves text reading in patients with hemianopic alexia: a controlled trial. *Neurology* 68, 1922–1930. doi: 10.1212/01.wnl.0000264002.30134.2a

Stanovich, K. E., Nathan, R. G., West, R. F., and Vala-Rossi, M. (1985). Children's word recognition in context: spreading activation, expectancy, and modularity. *Child Dev.* 56, 1418–1428. doi: 10.2307/1130461

Stifter, E., König, F., Lang, T., Bauer, P., Richter-Müksch, S., Velikay-Parel, M., et al. (2004). Reliability of a standardized reading chart system: variance component analysis, test-retest and inter-chart reliability. *Graefes Arch. Clin. Exp. Ophthalmol.* 242, 31–39. doi: 10.1007/s00417-003-0776-8

Stöhr, M., Dekowski, D., Bechrakis, N., Oeverhaus, M., and Eckstein, A. (2024). Evaluation of a retinal projection laser eyewear in subjects with visual impairment caused by corneal diseases in a randomized trial. *Ophthalmology* 131:545–556. doi: 10.1016/j.ophtha.2023.11.011

Tilanus, E. A. T., Segers, E., and Verhoeven, L. (2019). Predicting responsiveness to a sustained reading and spelling intervention in children with dyslexia. *Dyslexia* 25, 190–206. doi: 10.1002/dys.1614

Trauzettel-Klosinski, S., Dietz, K., and IReST Study Group (2012). Standardized assessment of reading performance: the New International Reading Speed Texts IReST. *Invest. Ophthalmol. Vis. Sci.* 53, 5452–5461. doi: 10.1167/iovs.11-8284

Wilkins, A. (2002). Coloured overlays and their effects on reading speed: a review. *Ophthal. Physiol. Opt.* 22, 448–454. doi: 10.1046/j.1475-1313.2002.00079.x

Wilkins, A., Sihra, N., and Smith, I. N. (2005). How precise do precision tints have to be and how many are necessary? *Ophthal. Physiol. Opt*. 25, 269–276. doi: 10.1111/j.1475-1313.2005.00279.x

Wilkins, A. J., Jeanes, R. J., Pumfrey, P. D., and Laskier, M. (1996). Rate of Reading Test®: its reliability, and its validity in the assessment of the effects of coloured overlays. *Ophthal. Physiol. Opt*. 16, 491–497. doi: 10.1046/j.1475-1313.1996.96000282.x

Zeri, F., Naroo, S. A., Zoccolotti, P., and De Luca, M. (2018). Pattern of reading eye movements during monovision contact lens wear in presbyopes. *Sci. Rep*. 8:15574. doi: 10.1038/s41598-018-33934-6

Zeri, F., Tavazzi, S., Punzi, M., Miglio, F., Evans, B. J. W., and De Luca, M. (2023). New Italian version of the Wilkins Rate of Reading Test: materials for repeated-measure designs in optometry and neuropsychological research. *Ophthal. Physiol. Opt*. 43, 629–639. doi: 10.1111/opo.13134

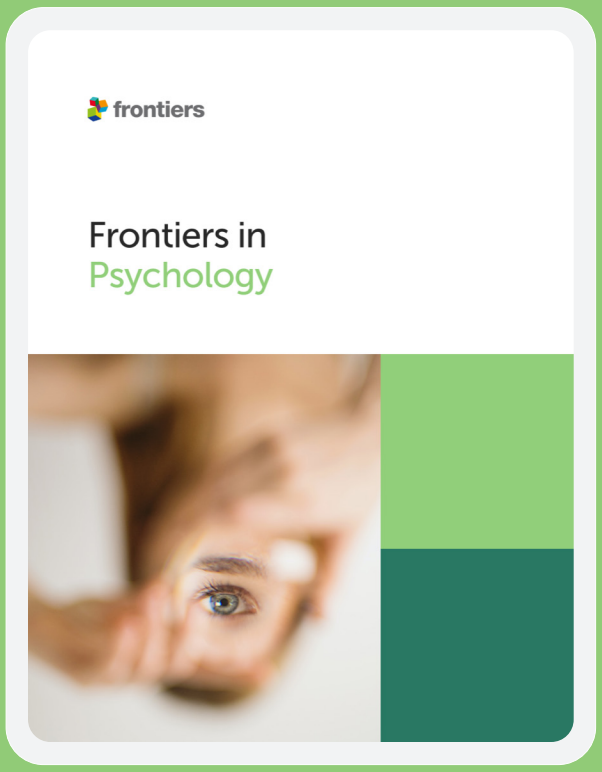
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact