



Virtual Reality and Serious Videogame-Based Instruments for Assessing Spatial Navigation in Alzheimer's Disease: A Systematic Review of Psychometric Properties

Juan Pablo Sánchez-Escudero¹ · Ana María Galvis-Herrera² · David Sánchez-Trujillo² ·
Laura Cristina Torres-López² · Cole J. Kennedy³ · Daniel Camilo Aguirre-Acevedo⁴ ·
Mauricio A. Garcia-Barrera³ · Natalia Trujillo^{1,5,6}

Received: 3 July 2023 / Accepted: 26 January 2024
© The Author(s) 2024

Abstract

Over the past decade, research using virtual reality and serious game-based instruments for assessing spatial navigation and spatial memory in at-risk and AD populations has risen. We systematically reviewed the literature since 2012 to identify and evaluate the methodological quality and risk of bias in the analyses of the psychometric properties of VRSG-based instruments. The search was conducted primarily in July–December 2022 and updated in November 2023 in eight major databases. The quality of instrument development and study design were analyzed in all studies. Measurement properties were defined and analyzed according to COSMIN guidelines. A total of 1078 unique records were screened, and following selection criteria, thirty-seven studies were analyzed. From these studies, 30 instruments were identified. Construct and criterion validity were the most reported measurement properties, while structural validity and internal consistency evidence were the least reported. Nineteen studies were deemed very good in construct validity, whereas 11 studies reporting diagnostic accuracy were deemed very good in quality. Limitations regarding theoretical framework and research design requirements were found in most of the studies. VRSG-based instruments are valuable additions to the current diagnostic toolkit for AD. Further research is required to establish the psychometric performance and clinical utility of VRSG-based instruments, particularly the instrument development, content validity, and diagnostic accuracy for preclinical AD screening scenarios. This review provides a straightforward synthesis of the state of the art of VRSG-based instruments and suggests future directions for research.

Keywords Alzheimer disease · Psychometric properties · COSMIN · Virtual reality · Systematic review · Serious games · Digital neuropsychology

✉ Juan Pablo Sánchez-Escudero
juanp.sanchez@udea.edu.co

Ana María Galvis-Herrera
ana.galvish@udea.edu.co

David Sánchez-Trujillo
david.sanchezt@udea.edu.co

Laura Cristina Torres-López
laura.torresl@udea.edu.co

Cole J. Kennedy
colekennedy@uvic.ca

Daniel Camilo Aguirre-Acevedo
daniel.aguirre@udea.edu.co

Mauricio A. Garcia-Barrera
mgarcia@uvic.ca

Natalia Trujillo
natalia.trujillo@udea.edu.co

¹ National College of Public Health, University of Antioquia, Antioquia, Colombia

² Department of Psychology, University of Antioquia, Antioquia, Colombia

³ Department of Psychology & Institute on Aging and Lifelong Health, University of Victoria, Victoria, BC, Canada

⁴ School of Medicine, University of Antioquia, Antioquia, Colombia

⁵ Atlantic Fellowship in Equity in Brain Health, Global Brain Health Institute, University of California, San Francisco, CA, USA

⁶ Stempel College of Public Health and Social Work, Florida International University, Miami, FL, USA

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease and the most common cause of dementia syndrome worldwide (Alzheimer's Association, 2023). Neuropathological hallmark AD signs include neurofibrillary tangles, senile plaques, and neuronal loss. These changes are primarily due to the accumulation of abnormal proteins such as beta-amyloid and phosphorylated tau (Alzheimer's Association, 2023; Jack et al., 2018). The progression of AD pathology is paralleled by the progression of cognitive symptoms, leading to a gradual functional decline and culminating in a clinical-pathological entity known as dementia by AD or Alzheimer's dementia (Alzheimer's Association, 2023). Given that the older age is one of the most significant risk factors for AD (Alzheimer's Association, 2023), the rising median age of the global population due to shifts in birth and death rates and the consistent rise in life expectancy is leading to an increase in AD prevalence (Alzheimer's Association, 2023; Levy et al., 2016). In North America alone, it is estimated that by 2050, 12.7 million individuals aged 65 years and over will be affected by the disease (Alzheimer's Association, 2023). As the prevalence and incidence of AD continue to rise, the development of valid and reliable instruments for cognitive assessment and follow-up from the preclinical stage of the disease is crucial (Levy et al., 2016).

According to the updated National Institute on Aging and Alzheimer's Association guidelines, AD is recognized as a continuum consisting of stages that reflect the progression of neurodegeneration and cognitive symptoms (Jack et al., 2018). While the presence of AD can be only determined by the presence of biomarkers (i.e., A β protein and tau protein), cognitive symptoms can be used to determine the staging of the disease (Jack et al., 2018). Thus, traditionally AD has been divided into three different stages: (1) the preclinical stage, also named cognitively unimpaired, where the neuropathological process has begun but no behavioral and cognitive symptoms can be detected; (2) the prodromal or early-stage symptomatic AD, often called mild cognitive impairment (MCI), where some cognitive changes can be detected with no interference in the daily level of independence; and the (3) dementia stage, where neurocognitive symptoms interfere with daily living activities (Jack et al., 2018; Sperling et al., 2011).

Despite the advantages in the detection of the preclinical stage of AD using biomarkers, the secondary effects and costs associated with the techniques used for their collection (e.g., lumbar puncture, positron emission tomography scans, functional magnetic resonance imaging) have promoted the search and identification of low cost, reduced secondary effect markers with acceptable

diagnostic accuracy that help in the screening process of the at-risk population (Levy et al., 2016; Lizio, 2020; Park, 2022; Sabbagh & Blennow, 2019).

In recent years, the development of cognitive markers based on information and communication technologies has been proposed as a suitable solution to this need (Ben-sadoun et al., 2018; Boot, 2015; Sacco et al., 2019; Sperling et al., 2014; Tong et al., 2016). Specifically in the context of preclinical detection of AD, spatial navigation and spatial memory task assessment using digital neuropsychology technologies, such as virtual reality and serious games, have been explored as alternatives for the detection of at-risk populations (Bayahya et al., 2021; Coughlan et al., 2018; Park, 2022; Poos et al., 2021; Puthusserypady et al., 2022).

Spatial Navigation and Spatial Memory in AD

Spatial navigation is the ability to determine and maintain a route from the starting point to the goal using different strategies and sources of information (Coughlan et al., 2018; Gazova et al., 2012). In complement, spatial memory is a cognitive function that allows an individual to recall the location of objects in the space and their spatial relations (Bird & Burgess, 2009; Jacobs, 2003; Kolarik & Ekstrom, 2015). Together, spatial navigation and memory are crucial for building and using effectively a unified mental representation of environment known as cognitive maps (Burgess, 2006; Epstein et al., 2017). Both in humans and other mobile species, the encoding and organization of spatial information is described using two frames of reference: allocentric and egocentric (Iachini et al., 2009). The egocentric representation of the environment involves using self-position and self-motion cues relative to the environment to codify the individuals' position and to set directions and distance to the goal (Coughlan et al., 2018; Iachini et al., 2009). In contrast, allocentric representations are independent of the self-position and do not change as the individual moves through space, relying on the landmarks in relation to each other instead to navigate (Iachini et al., 2009; Lester et al., 2017).

Neural correlates underlying spatial navigation and memory encompass several brain areas, forming a complex network for processing spatial information (Coughlan et al., 2018; Iachini et al., 2009; Jacobs, 2003). Structures in the medial temporal lobe, such as the hippocampus, entorhinal cortex, and parahippocampal cortex, support the cognitive map and are crucial for spatial navigation using allocentric representations (Colombo et al., 2017; Coughlan et al., 2018; Eichenbaum & Cohen, 2014). In contrast, egocentric spatial representations are primarily linked to the parietal lobe, particularly the medial and posterior parietal cortex, posterior cingulate, precuneus, and retrosplenial cortex (Coughlan et al., 2018). While traditional views have emphasized allocentric

navigation as hippocampal-dependent, contemporary models highlight the role of the hippocampus and parahippocampal cortex in egocentric navigation and acknowledge the contributions of other structures in cognitive map formation (Eichenbaum & Cohen, 2014; Goodroe et al., 2018; Kunz et al., 2021). Effective navigation in daily life often requires the integration of multiple sources of information, including both allocentric and egocentric representations requiring engagement from the frontal lobes, caudate nucleus, and thalamus in addition to medial temporal lobe structures (Coughlan et al., 2018; Morganti et al., 2013).

Recent research has linked deficits in spatial navigation and spatial memory tasks to the widespread neurodegeneration in medial temporal, parietal, and frontal brain regions during AD progression (Coughlan et al., 2018; Jacobs, 2003; Laczó et al., 2018; Nedelska et al., 2012). Due to AD pathophysiology effects on areas such as the entorhinal cortex, hippocampus, posterior cingulate cortex, and precuneus (Coughlan et al., 2018, 2019), individuals with AD and MCI due to AD present deficits in tasks involving allocentric and egocentric navigation strategies (Coughlan et al., 2019; Laczó et al., 2014, 2018; Vlcek & Laczó, 2014) and spatial memory (Kessels et al., 2015; Mitolo et al., 2013). Notably, individuals with AD dementia and MCI due to AD exhibit impairments in egocentric strategies, a pattern not observed in the preclinical stages of AD where allocentric performance is severely impaired (Laczó et al., 2018; Nedelska et al., 2012). This difference in the progression of allocentric and egocentric representations has relevant conceptual and practical implications for developing valid and reliable instruments and tasks to detect AD at-risk populations as it states the basis for hypothesis specification and construct validation of instruments (Ritchie et al., 2018; Ruggiero et al., 2018).

Tasks for Assessing Spatial Navigation in Humans

Since the concept of the cognitive map was first proposed by Tolman (1948) based on his work with rodents, researchers have adapted and designed different tasks to assess spatial navigation in humans in highly controlled settings such as laboratories (Burgess et al., 2004; Fernandez-Baizan et al., 2020a, 2020b; Schöberl et al., 2020). Paradigms such as the hidden goal task, a human version of the Morris water mazes for mice, have been successfully used for the assessment of allocentric and egocentric spatial navigation strategies, showing correlation with the brain structures implied in the spatial processing (Laczó et al., 2012; Nedelska et al., 2012).

Traditionally, tasks for assessing the egocentric and allocentric spatial memory adapt environments such as rooms or corridors to test specific hypotheses regarding spatial navigation strategies (Burgess et al., 2004; Fernandez-Baizan et al., 2019; Motes et al., 2006; Ribordy et al., 2013). While these environments provide high control of the experimental

situation, they can have limited ecologic validity and are expensive and challenging to reproduce in research and clinical settings (Campbell et al., 2009; Vasser et al., 2017). To address these limitations, researchers over the last two decades have increasingly turned to digital technologies, such as virtual reality (Castegnaro et al., 2022; Diersch & Wolbers, 2019; Nguyen et al., 2023; Puthusserypady et al., 2022; Ventura et al., 2013; Wiener et al., 2020) and serious games (Coughlan et al., 2019, 2020; Puthusserypady et al., 2022), to assess human spatial navigation skills.

Virtual reality consists in the use of interactive computer-generated graphics for simulating realistic environments, creating a sensory experience with different immersion degrees where the user perceives the world through screen or head-mounted devices (Pan et al., 2006; Zheng et al., 1998). On the other hand, serious games are developed with purposes other than entertainment (Deterding et al., 2011). Recent advancements in digital technologies, access to virtual reality devices, 3D modeling software, and game engines have led researchers to design and adapt their own navigational tasks and protocols (Alsbury-Nealy et al., 2021; Laczó et al., 2021; Vasser et al., 2017; Wiener et al., 2020). Among the different adaptations, virtual scenarios can incorporate game elements, such as badges, leaderboards, avatars, and rewarding systems, adaptations which have resulted in the creation of virtual reality-based serious games (Manera et al., 2017; Tong et al., 2014, 2016).

Virtual reality is a flexible technology that enables the creation of diverse environments without or with few constraints (e.g., device memory and processor capacity can limit the number and quality of objects to render in a scene). In the context of spatial navigation and memory assessment, guided by non-human research, different types of mazes have been adapted for assessing allocentric and egocentric spatial navigation strategies (Lee et al., 2014; Nguyen et al., 2023). More recently, tasks involving route learning, path integration, wayfinding, and landmark placement have been adapted to virtual reality scenarios (Puthusserypady et al., 2022; Wiener et al., 2020).

Although virtual reality and serious game-based (VRSG-based) instruments are developed following recommendations for software and experimental psychological task programming, aspects related to the quality of design and analysis of psychometric properties are not usually assessed using systematic criteria (Silva et al., 2023). Since test reliability and validity of the scores are prerequisites for robust assessment, analyzing the methodological quality of the studies on measurement properties is crucial for further advances in the field (Mokkink et al., 2018). In recent years, some other reviews have focused on the contrast between real-world and virtual reality tasks (Cogné et al., 2017; Tuena et al., 2021) or summarize the diagnostic accuracy for detecting mild cognitive impairment (MCI) and dementia (Chan

et al., 2021; Molina da Costa et al., 2020). However, the systematic assessment of the methodological quality of studies on psychometric properties remains unattended so far.

Therefore, the specific objectives of this review were (1) identify and describe the existing VRSG-based instruments developed for assessing spatial navigation and spatial memory in samples with preclinical AD, MCI due to AD, and dementia by AD; (2) evaluate the methodological quality of the studies; and (3) establish the risk of bias in the analysis of psychometric properties of the instruments used in the studies identified using COSMIN guidelines. This review will provide information regarding the quality of existing evidence, offering researchers and clinical neuropsychologists a comprehensive analysis of state of the art regarding VRSG-based instruments for the neuropsychological assessment of spatial navigation and spatial memory in at-risk populations for AD. Based on the analyzed evidence, recommendations regarding using these instruments in clinical and research settings and future directions for advances in digital neuropsychology research are provided.

Methods

A protocol was developed and registered with PROSPERO (CRD42022339039) in accordance with COSMIN guidelines, which established a procedure for the quality appraisal of instruments (Mokkink et al., 2018). Although the COSMIN methodology is originally focused on Patient-Reported Outcome Measures usage, it provides valuable and standardized criteria for evaluating other types of measurement instruments (Mokkink et al., 2018). Key components of COSMIN include a detailed checklist for assessing the methodological quality of studies for each measurement property and an emphasis on evaluating the generalizability and relevance of the study findings (Mokkink et al., 2018).

Data Sources and Searches

A systematic literature search, initially conducted from July to November 2022 and subsequently updated in November 2023, aimed to identify published articles examining the psychometric properties of VRSG-based instruments for assessing spatial memory and navigation in AD. Eight databases were selected for sourcing peer-reviewed literature: Scopus, PubMed/Medline, EBSCO, APA PsycINFO, Web of Science, EMBASE, SciELO, and RedALyC. A supplementary gray literature review was implemented to encompass a broader range of relevant instruments. The extended search included ProQuest and LILACS for theses and dissertations, as well as PsycArticles, Google Scholar, WHO IRIS, Index Psicología-Periódicos, SciELO Preprints, and Open Science Framework for other gray literature sources.

The results from these sources were subjected to the same rigorous analysis and eligibility criteria as those identified from peer-reviewed literature databases. These efforts aimed to capture literature in the early stages or in preparation for peer review and to reduce the influence of publication bias on our results (Winters & Weir, 2017). The search strategy also involved Spanish-language terms, which were identified through an initial search in the library systems and repositories of universities offering graduate programs in neuropsychology and related disciplines. See supplementary material for a detailed description of search equations.

Study Selection

Studies met inclusion criteria if they (1) report original data on the analysis of psychometric properties of VRSG-based instruments for the assessment of spatial navigation or spatial memory; (2) involve participants with an increased risk for AD (i.e., genetic risk for sporadic AD, suspected MCI due to AD) or participants in the continuum of AD; (3) use gamification, serious games, or virtual reality to develop or adapt the instruments; (4) have accessibility of the full text; (5) are published in English or Spanish language, regardless of the original language used for developing the instrument; and (6) are published between 2012 and 2023. The date range was chosen to reflect the growing utilization of virtual reality-based assessments and technological advancements over the past decade and the recent developments in virtual reality technology and 3D modeling software on the regard of neuropsychological assessment (Cipresso et al., 2018; Diersch & Wolbers, 2019; Krohn et al., 2020).

Studies identified were entered into the Rayyan online reviewing system (Ouzzani et al., 2016) for screening. Then, the duplicated studies were removed. Four authors (J.P., A.G., D.S., and L.T.) screened titles and abstracts to determine whether the studies met the inclusion criteria. Each study was randomly assigned to a pair of independent reviewers for inclusion criteria assessment. Uncertainty about eligibility or data accuracy was resolved by consensus between reviewers. The criteria for exclusion described in Table 1 were used.

Data Extraction

Two independent reviewers extracted the relevant data for each study. Data extraction included study population characteristics (i.e., number of participants, age, gender, risk status for AD), publication details (i.e., authors, year of publication), instrument description, virtual reality or serious game technology, study design, neuropsychological correlates, and main findings on the psychometric properties (i.e., indicators for reliability such as internal consistency, test–retest, and indicators for content validity, criterion

Table 1 Criteria for the study exclusion

Criteria	Description
Language	Paper published in a language other than Spanish or English
Outcome	The primary outcome studied is not AD or MCI
Population	The participants are not diagnosed with AD or MCI
Publication type	The paper is not an empirical research paper
Study design	Randomized clinical trials, systematic and literature review were excluded unless they specifically addressed the assessment of psychometric properties
No VRSG-based instrument	The study used a software-based measure, but there is no use of gamification strategies or virtual reality scenarios
Instrument modality	The study did not use a software-based instrument (e.g., paper–pencil test and real environment assessments were used)
Construct	The study does not present evidence on spatial navigation or spatial memory; instead, it analyzes other cognitive processes
Previously reported findings	The findings of this study have been documented in another peer-reviewed publication

validity, and construct validity) extracted and exported in Microsoft Excel files following COSMIN guidelines (Prinsen et al., 2018).

Instrument development design and methodological quality of the included studies were evaluated separately in accordance with COSMIN guidelines (Prinsen et al., 2018). Psychometric properties (i.e., content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, and construct validity) were analyzed when the original data was available. All psychometric properties were defined according to COSMIN guidelines. See supplementary material for a detailed description of the definition of measurement properties.

This review used a comprehensive approach to assess the analysis of psychometric properties in the studies identified. Thus, the performance differences between groups with different levels of risk for AD were considered evidence of construct validity (i.e., known-group validity), while correlations between instrument scores and neuropsychological and neurological measures were seen as evidence of convergent validity. Diagnostic accuracy measures were evaluated according to COSMIN guidelines for criterion validity. This approach allowed the review to obtain relevant information about the measuring properties, even if the authors did not explicitly present this information.

Assessment of Psychometric Properties

The methodological quality of studies on measurement properties of VRSG-based instruments was assessed by following the risk of bias checklist and the quality judgment criteria contained in the ten-item COSMIN checklist (Mokkink et al., 2018). For each article, two independent reviewers assessed the methodological quality of the studies and the risk of bias within the psychometric properties reported by the authors. The results were rated as “very

good,” “adequate,” “doubtful,” or “inadequate” according to the four-point rating system and the *worst score counts principle*, taking the lowest rating of any standard in the analysis of the psychometric property analyzed (Mokkink et al., 2018).

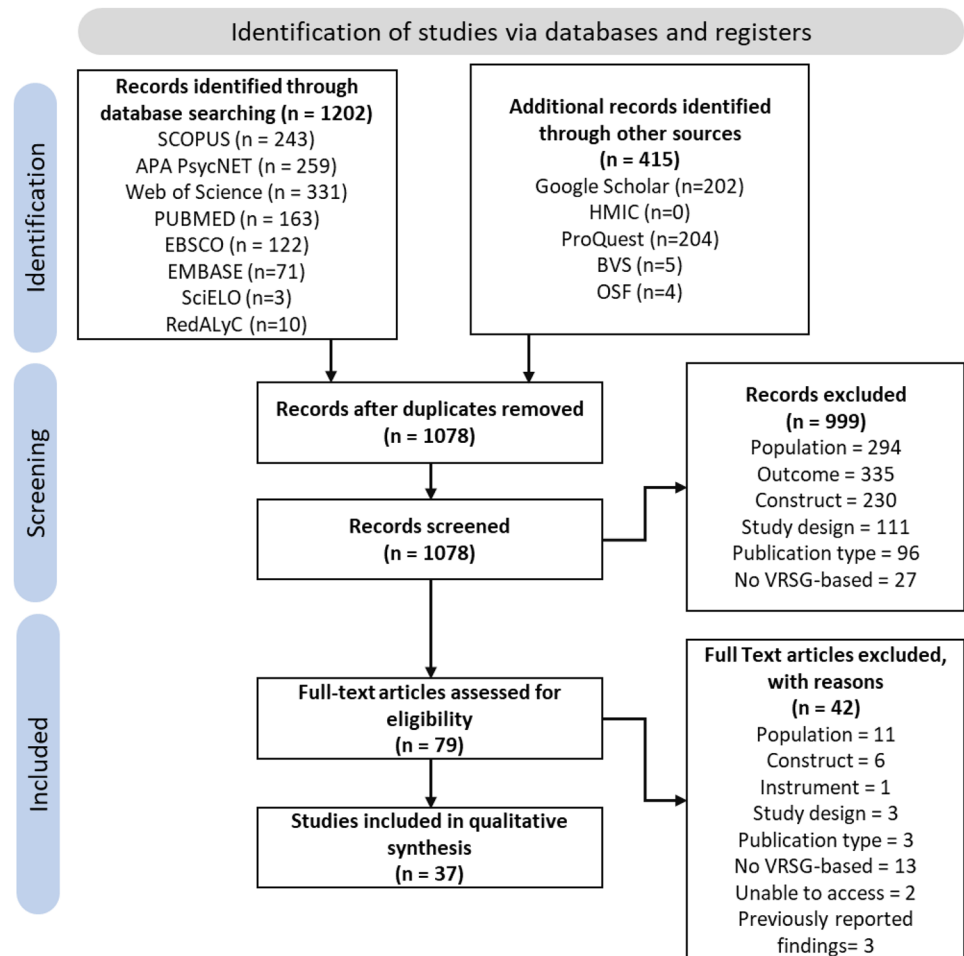
Results

The search strategy returned 1078 unique records, from which 79 were selected for full-text screening. Of these, 37 articles reported evidence about the psychometric properties of 30 instruments. As three of the included studies presented evidence on more than one instrument (Coughlan et al., 2020; Morganti et al., 2013; Puthusseryppady et al., 2022), COSMIN criteria were assessed in each VRSG instrument. Figure 1 shows the results of the search in the databases. Due to the eligibility criteria, pilot studies using healthy populations for developing the instrument were excluded. All studies identified described virtual reality-based instruments. Among these studies, ten of them included gamification strategies (Bayahya et al., 2021; Bierbrauer et al., 2020; Colmant et al., 2023; Coughlan et al., 2019, 2020; Gellersen et al., 2021; Lee et al., 2014; Pink et al., 2023; Puthusseryppady et al., 2022; Tarnanas et al., 2015).

Study Summary and Assessment Methodology

A summary of the findings is presented in Table 2. Among the VRSG-based instruments identified, 13 (43.3%) were developed in English, while four instruments (13.3%) were developed in Italian. Additionally, two instruments (6.7%) each were developed in Czech, French, Greek, and Korean. At least one instrument (3.3%) was developed in each of Farsi, German, Portuguese, and Spanish, demonstrating the global diversity of these tools. Regarding the technology used, four instruments (13.3%) were applied on iPad devices

Fig. 1 PRISMA flow diagram.
Summary of search



(Coughlan et al., 2019, 2020; Puthusseryppady et al., 2022; Ritchie et al., 2018), whereas one was developed to be used on mobile phones (Tarnanas et al., 2015), and remaining instruments were developed to be applied using personal computers and VR devices. A description of the instruments is provided as supplementary material. Among the studies using VR, five used immersive modalities (Castegnaro et al., 2022; Da Costa et al., 2022; Moussavi et al., 2022; Tarnanas et al., 2015; Zen et al., 2013), where the examinee wears a head-mounted display to see while a sensor collects data about their movement and position (Morganti et al., 2013).

Psychometric Property Assessment and Level of Evidence

Overall COSMIN Assessment of Instruments The overall results of the COSMIN assessments are shown in Table 3. Among all the instruments identified, the psychometric properties reported were structural validity, internal consistency, criterion validity, and construct validity. None of the instruments reported evidence for all

these psychometric properties, revealing a gap in robust psychometric evidence for the measurement properties of instruments used across the literature. Regarding the general quality of the studies, three were deemed adequate or very good (Allison et al., 2016; Caffò et al., 2018; Coughlan et al., 2019). Given the innovative nature of VRSG-based instruments, the lack of pilot studies did not automatically lead to an inadequate or doubtful rating. Studies classified under these categories typically featured insufficient descriptions of their target populations, the constructs they intended to measure, or the theoretical frameworks underpinning these constructs (Mokkink et al., 2019).

Structural Validity Two studies (5.4%) reported evidence of structural validity (Allison et al., 2019; Bellassen et al., 2012). One of the studies (Bellassen et al., 2012) used exploratory factor analysis rather than confirmatory factor analysis required to be classified as very good according to the COSMIN guidelines (Mokkink et al., 2018). However, for both studies, no sample size calculation or rationale was presented, leading to concerns about the sample size adequacy for the analyses.

Table 2 Summary of studies

Reference	VRSG-based instrument	Participants	Technology				Mode of administration	Language	Country	
			Subgroups	n	Sex					
					Age	SD				
(%female)										
Bellassen et al. (2012)	The starmaze task	HC	20	28.2	5.7	55%	Virtual reality	Personal computer	French	France
		HC	19	48.3	6.8	47%				
		HC	24	72	4.2	54%				
		aMCI due to AD	14	70.8	8.8	29%				
		AD dementia	16	70	9.0	50%				
Caffò et al. (2012)	Virtual reorientation test (VReoT)	FTD-FV	11	68.2	7.8	46%	Virtual reality	Personal computer	Italian	Italy
		aMCI-SD	28	69.9	5.2	60%				
		aMCI-MD	23	73.9	4.7	65%				
		HC	53	68.1	6.0	55%				
		aMCI	25	78	4.7	72%				
Tarnanas et al. (2012)	Virtual museum system (VAP-M)	HC	25	77	4.9	80%	Virtual reality	Personal computer	Greek	Greece
		HC	26	77.2	5.3	50%				
		AD dementia	26	81.0	6.3	50%				
		HC	26	77.2	5.3	50%				
		AD dementia	26	81.0	6.3	50%				
Morganti et al. (2013)	VR-maze spatial task (VRMST)	Probable AD dementia	8	71.1	11.8	50%	Virtual reality	Personal computer	English	Canada
		HC	11	70.4	4.3	45%				
		HC	20	70.8	5.2	55%				
		Probable AD	20	72.4	5.6	50%				
		aMCI	20	70.7	5.0	45%				
Lesk et al. (2014)	Virtual reality for early detection of AD (VREAD)	HC	22	69.8	4.6	77%	Virtual reality	Personal computer	English	England
		MCI	9	72.7	6.6	44%				
		HC	20	70.8	5.2	55%				
		Probable AD dementia	20	72.4	5.6	50%				
		aMCI	20	70.7	5.0	45%				
Kunz et al. (2015)	Object location memory task	HC	38	22.8	0.5	51%	Virtual reality	N/R	English	USA
		ε3ε4 carriers	37	22.3	0.5	53%				
		Probable AD dementia	15	83	5.6	27%				
		aMCI	15	78	5.5	27%				
		HC	15	74	7.4	40%				
Serino et al. (2015)	Virtual room environment	HC	76	70.1	13.3	65%	Virtual reality	Personal computer	Italian	Italy
		aMCI	85	72.6	10.1	62%				
		Early AD dementia	86	76.6	10.6	63%				
		HC	76	70.1	13.3	65%				
		aMCI	85	72.6	10.1	62%				
Tarnanas et al. (2015)	Complex activities of daily living	HC	76	70.1	13.3	65%	Serious game	Mobile phone	Greek	Greece
		aMCI	85	72.6	10.1	62%				
		Early AD dementia	86	76.6	10.6	63%				
		HC	76	70.1	13.3	65%				
		aMCI	85	72.6	10.1	62%				

Table 2 (continued)

Reference	VRSG-based instrument	Participants	Subgroups				Technology	Mode of administration	Language	Country
			n	Age		Sex (%female)				
				Mean	SD					
Allison et al. (2016)	Cognitive mapping task	HC biomarker +	12	73.0	9	58%	Virtual reality	Personal computer	English	USA
		HC biomarker –	42	69.0	9	55%				
	Platform task	Early AD dementia	15	77.0	10	40%	Virtual reality	Personal computer	English	England
		HC	10	70.3	6.5	40%				
Migo et al. (2016)	Virtual reorientation test (VReoT)	aMCI	8	69.6	5.8	38%	Virtual reality	Personal computer	Italian	Italy
		HC	203	68.6	7.1	47%				
		aMCI SD	14	70.9	8	71%				
		aMCI MD	22	73	4.7	50%				
Konishi et al. (2018)	Concurrent spatial discrimination task (CSDLT)	aMCI MD	16	74.9	8	56%	Virtual reality	Personal computer	English	Canada
		naMCI MD	10	75.3	6.7	90%				
		Probable AD dementia	21	74.3	6.6	67%				
		e3e4 carriers	15	65.6	4.5	N/R				
		HC	49	66.5	4.6	N/R				
		HC	30	69.8	1.4	50%				
		AD dementia	20	73.6	2.5	65%				
Mohammadi et al. (2018)	Virtual neighborhood task (VRNT)	aMCI SD	30	70.0	1.6	63%	Virtual reality	Personal computer	Farsi	Iran
		aMCI MD	30	70.0	1.7	60%				
		HC	20	67.5	7.1	10%				
		aMCI due to AD	28	74.5	5.9	64%				
Parizkova et al. (2018)	Y-maze strategy assessment (yVSA)	AD dementia	21	73.2	6.9	57%	Virtual reality	Personal computer	Czech	Czech Republic
		Non-AD FH	81	54.2	10.8	70%				
		AD FH	94	55.3	6.6	71%				
		Probable AD dementia	15	86	4.0	73%				
Ritchie et al. (2018)	Reality supermarket trolley task	HC	15	83	5.6	80%	Virtual reality	Personal computer	Italian	Italy
		ADRC	30	73	4	43%				
		RPR	60	67	5	68%				
		e3e3 carriers	29	62.5	5.3	48%				
Serino et al. (2018)	Virtual room environment	e3e4 carriers	31	62.3	6.2	64%	Virtual reality	Tablet	English	England
		Benchmark players	27,108	58.9	5.9	51%				
		e3e3 carriers	76	42.6	2.4	47.30%				
		e3e4 carriers	23	46.1	4.2	52.10%				
Allison et al. (2019)	Modified cognitive mapping task						Virtual reality	Personal computer	English	USA
Coughlan et al. (2019)	Sea Hero Quest (SHQ)						Virtual reality	Tablet	English	England
Bierbrauer et al. (2020)	The apple game						Virtual reality	Personal computer	English	Belgium

Table 2 (continued)

Reference	VRSG-based instrument	Participants	Subgroups				Technology	Mode of administration	Language	Country				
			n	Age	Sex									
					Mean	SD					(%female)			
Coughlan et al. (2020)	Virtual supermarket test (VST)	ε3ε3 carriers	33	64.4	5.6	44%	Virtual reality	Tablet	English	England				
		ε3ε4 carriers	31	63.7	6.4	70%								
	Sea Hero Quest (SHQ)	ε3ε3 carriers	33	64.4	5.6	44%					Virtual reality	Tablet	English	England
		ε3ε4 carriers	31	63.7	6.4	70%								
Davis and Sikorskii (2020)	VR simulation of a large senior residence	Probable AD dementia	7	76.6	5.0	47%	Virtual reality	Personal computer	English	USA				
	HC	8	75	1.2	50%									
Levine et al. (2020)	Cognitive mapping task	No progression	78	70.6	7.9	50%	Virtual reality	Personal computer	English	USA				
		Progression	17	78.9	7.5	41%								
Bayahya et al. (2021)	The virtual scenario	Control	65	60	7.9	85%	Virtual reality	Personal computer	N/R	Saudi Arabia				
		MCI	20	68.5	9.3	60%								
		Dementia	30	80	8.1	33%								
		ε3ε3 carriers	26	63.4	6.1	50%								
Gellersen et al. (2021)	Sea Hero Quest (SHQ)	ε3ε4 carriers	20	64.8	6.8	25%	Virtual reality	Tablet	English	England				
		ε3ε4 carriers	3	63.3	0.6	33%								
		HC	78	68.2	6.8	72%								
		aMCI	75	71.5	7.4	75%								
Laczó et al. (2021)	Navigation test suite	Mild AD dementia	66	75.9	8.1	66%	Virtual reality	Personal computer	Czech	Czech Republic				
		Young controls	53	24.1	3	59%								
		HC	24	68.8	5.7	67%								
		aMCI	23	71.7	7.5	57%								
Castegnaro et al. (2022)	The object location task	aMCI	19	73	6.5	68%	Virtual reality	Personal computer	Portuguese	Brazil				
		HC	29	70	5.3	69%								
		HC	30	68.7	5.8	70%								
		Non-AD MCI	31	70.3	7.7	48%								
Da Costa et al. (2022)	SOIVET (maze task and route task)	aMCI due to AD	33	72.3	6.4	61%	Virtual reality	Personal computer	Czech	Czech Republic				
		Mild AD dementia	28	74.3	5.9	64%								
		Probable AD dementia	65	72.4	8.7	43%								
		MCI	20	72.9	7.9	35%								
Moussavi et al. (2022)	Virtual reality navigational experiment (VRN)	aMCI	36	75.3	6.2	61%	Virtual reality	Personal computer	English	Canada				
		HC	56	73.0	6.5	55%								
		HC	56	73.0	6.5	55%								
		HC	56	73.0	6.5	55%								
Park (2022)	Spatial cognitive task based on virtual reality (SCT-VR)	aMCI	36	75.3	6.2	61%	Virtual reality	Personal computer	Korean	Korea				
		HC	56	73.0	6.5	55%								
		HC	56	73.0	6.5	55%								
		HC	56	73.0	6.5	55%								

Table 2 (continued)

Reference	VRSG-based instrument	Participants		Technology			Mode of administration	Language	Country	
		Subgroups	n	Age	Sex					
					Mean	SD				
Puthusseryppady et al. (2022)	Virtual supermarket test (VST)	HC	21	68.4	7.6	56%	Virtual reality	Tablet	English	England
		Probable AD dementia	16	70.3	6.6	50%				
	Sea Hero Quest (SHQ)	HC	21	68.4	7.6	56%	Virtual reality	Tablet	English	England
		Probable AD dementia	16	70.3	6.6	50%				
Silva et al. (2023)	SOIVET (maze task and route task)	HC	24	71.3	5.6	63%	Virtual reality	Personal computer	Portuguese	Brazil
		MCI	19	72.3	5.6	63%				
	The apple game	e4 carriers	73	58.6	15.2	62%	Serious game	Personal computer	English	Belgium
Colmant et al. (2023)		e4 noncarriers	198	57.3	16.3	65%				
	Virtual reality-based path integration task	e3e4 carriers	31	43.1	17.1	64%	Virtual reality	Personal computer	German	Germany
		Controls	73	45.3	15.9	76%				
Pink et al. (2023)	Virtual Morris water navigation (VMWN)	MCI due to AD	9	76.7	6.2	78%	Virtual reality	Personal computer	Spanish	Chile
		HC	9	71.2	8.5	56%				

HC healthy control, AD Alzheimer's disease, MCI mild cognitive impairment, aMCI amnesic mild cognitive impairment, aMCI-SD amnesic mild cognitive impairment single domain, aMCI-MD amnesic mild cognitive impairment multiple domain, FTD-FV frontotemporal dementia frontal variant, naMCI non-amnesic mild cognitive impairment, N/R not reported, FH familial history, ADRC sample from the Alzheimer Disease Research Center, RPR sample from the Washington University in St. Louis Research Participant Registry

Table 3 COSMIN assessment of instruments

Reference	VRSG-based instrument	General quality	Structural validity	Internal consistency	Test–retest reliability	Criterion validity	Convergent validity	Known-group validity
Bellassen et al. (2012)	The starmaze task	Inadequate	Adequate	NA	NA	Doubtful	NA	Adequate
Caffò et al. (2012)	Virtual reorientation test	Doubtful	NA	NA	NA	Very good	NA	Adequate
Tarnanas et al. (2012)	Virtual museum system	Inadequate	NA	NA	NA	NA	NA	Very good
Morganti et al. (2013)	VR-maze spatial task	Doubtful	NA	NA	NA	NA	Very good	Very good
	VR-road map task	Doubtful	NA	NA	NA	NA	Adequate	Adequate
Zen et al. (2013)	Virtual reality navigational experiment	Doubtful	NA	NA	NA	NA	NA	Doubtful
Lesk et al. (2014)	Virtual reality for early detection of AD	Inadequate	NA	NA	NA	NA	Doubtful	Doubtful
Lee et al. (2014)	VRAM task	Inadequate	NA	NA	NA	NA	Adequate	NA
Kunz et al. (2015)	Object location memory task	Inadequate	NA	NA	NA	NA	Very good	Very good
Serino et al. (2015)	Virtual room environment	Doubtful	NA	NA	NA	NA	NA	Very good
Tarnanas et al. (2015)	Complex activities of daily living	Doubtful	NA	NA	Adequate	NA	NA	Very good
Allison et al. (2016)	Cognitive mapping task	Adequate	NA	Very good	NA	Very good	NA	Very good
Migo et al. (2016)	Platform task	Doubtful	NA	NA	NA	NA	NA	Adequate
Caffò et al. (2018)	Virtual reorientation test	Adequate	NA	Very good	NA	NA	NA	Adequate
Konishi et al. (2018)	Concurrent spatial discrimination task	Doubtful	NA	NA	NA	NA	NA	Adequate
Mohammadi et al. (2018)	Virtual neighborhood task	Doubtful	NA	NA	NA	NA	NA	Doubtful
Parizkova et al. (2018)	Y-maze strategy assessment	Doubtful	NA	NA	NA	NA	Very good	Very good
Ritchie et al. (2018)	Reality supermarket trolley task	Inadequate	NA	NA	NA	NA	Very good	Doubtful
Serino et al. (2018)	Virtual reality-based procedure	Doubtful	NA	NA	NA	NA	NA	Doubtful
Allison et al. (2019)	Modified cognitive mapping task	Doubtful	Adequate	Very good	Adequate	Very good	NA	NA
Coughlan et al. (2019)	Sea Hero Quest	Very good	NA	NA	NA	Very good	NA	Very good
Bierbrauer et al. (2020)	The apple game	Adequate	NA	NA	NA	NA	Very good	Very good

Table 3 (continued)

Reference	VRSG-based instrument	General quality	Structural validity	Internal consistency	Test–retest reliability	Criterion validity	Convergent validity	Known-group validity
Coughlan et al. (2020)	Virtual super-market test	Inadequate	NA	NA	Doubtful	NA	NA	NA
	Sea Hero Quest	Inadequate	NA	NA	Doubtful	NA	NA	NA
Davis and Sikorskii (2020)	VR simulation of a large senior residence	Inadequate	NA	NA	NA	NA	NA	Doubtful
Levine et al. (2020)	Cognitive mapping task	Inadequate	NA	NA	NA	Very good	NA	NA
	Route learning task	Inadequate	NA	NA	NA	Very good	NA	NA
Bayahya et al. (2021)	The virtual scenario	Inadequate	NA	NA	NA	NA	NA	Doubtful
Gellersen et al. (2021)	Sea Hero Quest	Inadequate	NA	NA	NA	NA	NA	Very good
Laczó et al. (2021)	Navigation test suite	Inadequate	NA	NA	NA	NA	NA	Very good
Castegnaro et al. (2022)	The object location task	Inadequate	NA	NA	NA	Very good	Very good	Very good
Da Costa et al. (2022)	Maze task and route task (SOIVET)	Doubtful	NA	NA	NA	Very good	Doubtful	Adequate
Laczó et al. (2022)	Navigation test suite	Doubtful	NA	NA	NA	Very good	Very good	Very good
Moussavi et al. (2022)	Virtual reality navigational experiment	Doubtful	NA	NA	NA	NA	NA	Doubtful
Park (2022)	Spatial cognitive task based on virtual reality (SCT-VR)	Inadequate	NA	NA	Adequate	Very good	NA	Adequate
Puthusserypady et al. (2022)	Virtual super-market test	Inadequate	NA	NA	NA	NA	Adequate	Very good
	Sea Hero Quest	Inadequate	NA	NA	NA	NA	Adequate	Very good
Silva et al. (2023)	Maze task and route task (SOIVET)	Adequate	NA	NA	Adequate	NA	NA	NA
Colmant et al. (2023)	The apple game	Adequate	NA	NA	NA	NA	NA	Very good
Pink et al. (2023)	Virtual reality-based path integration task	Adequate	NA	NA	NA	NA	NA	Doubtful
Plaza-Rosales et al. (2023)	Virtual Morris water navigation	Doubtful	NA	NA	NA	NA	NA	Very good

NA not assessed

Internal Consistency Three studies (8.1%) reporting the internal consistency of two instruments were identified (Allison et al., 2016, 2019; Caffò et al., 2018). The internal consistency for the *virtual reorientation test* was 0.79

(Caffò et al., 2018); Cronbach's alpha was calculated using the subtask score as the units of measurement. Additionally, the internal consistency of the *cognitive mapping task* was reported differently between the studies (Allison et al.,

2016, 2019). Both studies reported on the internal consistency of the wayfinding task, which comprises a learning phase, where participants explored a virtual environment and placed an “X” at all landmark locations on a blank 2D map, and a retrieval phase, where participants completed a series of supplementary tasks (i.e., landmark free recall, landmark location memory, and landmark identification memory) (Allison et al., 2016, 2019).

In the first study (Allison et al., 2016), Cronbach’s alpha was calculated using scores from the supplementary tasks, resulting in a reported consistency of $\alpha=0.86$. For the modified version (Allison et al., 2019), the internal consistency was calculated separately. For continuous scores, Cronbach’s alpha was used, resulting in the learning phase ($\alpha=0.83$) being reported as consistent, in contrast to retrieval phase ($\alpha=0.35$) which presented a low consistency. Kuder Richardson-20 was used for dichotomic scores, resulting in different levels of consistency for the landmark identification ($KR_{20}=0.87$), scene recognition ($KR_{20}=0.87$), and free recall ($KR_{20}=0.62$) subtasks (Allison et al., 2019).

Reliability (Test–Retest) Four studies (10.8%) reporting evidence about the test–retest reliability of five instruments were identified (Allison et al., 2019; Coughlan et al., 2020; Park, 2022; Tarnanas et al., 2015). The quality of evidence was deemed as very good in one study (Tarnanas et al., 2015), adequate in another (Allison et al., 2019), and doubtful in the remaining two (Coughlan et al., 2020; Park, 2022). The recall period varied between 2 weeks (Park, 2022) and 18 months (Coughlan et al., 2020), with 3 months being the most frequent (Allison et al., 2019; Tarnanas et al., 2015). No rationale for the selection of the recall period was provided in any of the studies.

The test–retest reliability for the *spatial cognitive task* based on virtual reality (Park, 2022) was high ($ICC=0.982, p<0.001$). In the case of the *virtual supermarket* (Coughlan et al., 2020), the test–retest reliability was moderate to adequate, depending on the subtask. Thus, the reliability of the egocentric score was adequate ($ICC=0.72$), whereas the heading subtask was moderate ($ICC=0.50$). For the *Sea Hero Quest* (Coughlan et al., 2020), the test–retest reliability was moderate for both distance ($ICC=0.50$) and duration ($ICC=0.48$) scores.

The accuracy-based tasks for the *complex activities of daily living* (Tarnanas et al., 2015) showed fair to moderate reliability ($ICC=0.33–0.57$), whereas latency-based measures exhibited good to excellent reliability measures ($ICC=0.69–0.85$). The *modified cognitive mapping task* (Allison et al., 2019) showed adequate reliability. Specifically, the tasks learning phase ($ICC=0.719$), free recall ($ICC=0.570$), and landmark identification ($ICC=0.722$) exhibited good reliability. Effects of recall period on performance are yet to be determined.

Criterion Validity Ten studies (27%) reporting evidence about the criterion validity and diagnostic accuracy of 11 instruments were identified. Nine studies were deemed very good as they calculated the area under the ROC curve for the continuous scores. Seven studies determined the sensitivity and specificity of the instruments based on the cutoff score estimated by Youden’s index (Allison et al., 2016, 2019; Bellassen et al., 2012; Da Costa et al., 2022; Laczó et al., 2022; Levine et al., 2020; Park, 2022).

The studies showed a heterogeneous selection of the standard of reference. Clinical and neuropsychological assessments were used to determine the diagnostic status of the participants in nine studies (90%), while one study (Coughlan et al., 2019) used the APOE risk status determined by genotyping techniques as the standard of reference (see supplementary material for additional information regarding the diagnostic accuracy). Additionally, six studies employed biomarkers to complement the diagnosis (Allison et al., 2016, 2019; Castegnaro et al., 2022; Coughlan et al., 2019; Laczó et al., 2022; Levine et al., 2020).

Construct Validity (Convergent Validity) Nine studies (24.3%) reported data about the correlation between software-based instruments and neuropsychological or neurological measures, constituting evidence of convergent validity according to COSMIN guidelines (Castegnaro et al., 2022; Da Costa et al., 2022; Kunz et al., 2015; Laczó et al., 2022; Lee et al., 2014; Lesk et al., 2014; Morganti et al., 2013; Parizkova et al., 2018; Ritchie et al., 2018). Of the studies reported, eight were deemed as very good (Bierbrauer et al., 2020; Castegnaro et al., 2022; Laczó et al., 2022; Morganti et al., 2013; Parizkova et al., 2018; Ritchie et al., 2018), two as adequate (Lee et al., 2014; Morganti et al., 2013), and the three remaining as doubtful (Da Costa et al., 2022; Lesk et al., 2014).

Neurological correlates were reported in four studies (Bierbrauer et al., 2020; Kunz et al., 2015; Laczó et al., 2022; Parizkova et al., 2018). The brain activity related to structures implicated in the visuospatial processing as the grid cells in the entorhinal cortex (Bierbrauer et al., 2020; Kunz et al., 2015), as well as the hippocampal volume (Laczó et al., 2022; Parizkova et al., 2018), showed negative correlations with the performance on tasks from the *object location memory task* (Kunz et al., 2015), the *apple game* (Bierbrauer et al., 2020), the *yVSA* (Parizkova et al., 2018), and the *navigation test suite* (Laczó et al., 2022).

Neuropsychological correlates were reported in five studies (Da Costa et al., 2022; Lee et al., 2014; Lesk et al., 2014; Morganti et al., 2013; Ritchie et al., 2018). The global score of cognitive functioning scales was reported in four studies (Da Costa et al., 2022; Lesk et al., 2014; Morganti

et al., 2013; Ritchie et al., 2018). The *MMSE*, *Dementia Risk Score*, and *Addenbrooke's Cognitive Examination* were used in three studies (Da Costa et al., 2022; Lesk et al., 2014; Morganti et al., 2013). The *Corsi Test* scores (i.e., forward, backward, and supraspan) (Morganti et al., 2013), the *Manikin test* (Morganti et al., 2013), the *Rey Complex Figure Test* (Lee et al., 2014), the *Money Road-Map* test (Da Costa et al., 2022), and *Benton's Judgment of Line Orientation* test (Da Costa et al., 2022) were reported as visuospatial correlates. The *VR-maze spatial task* (Morganti et al., 2013), *VR-road map task* (Morganti et al., 2013), *VRAM task* (Lee et al., 2014), and the *SOIVET* task (Da Costa et al., 2022) showed a positive and statistically significant correlation with the visuospatial measures. Correlation coefficient values between VRSG-based instrument and neuropsychological tests are provided in supplementary material.

Construct Validity (Known-Group Validity) Based on previous literature (Fernandez-Baizan et al., 2020a, 2020b; Laczó et al., 2021; Levine et al., 2020; Nedelska et al., 2012; Parizkova et al., 2018), a set of a priori hypotheses were established to analyze the construct validity through known-group comparison. Thus, according to the natural progression of AD and previous findings (Allison et al., 2016; Colombo et al., 2017; Coughlan et al., 2018), it was expected that older participants, and those in a more advanced stage of the disease, would perform worse than younger participants and those in preclinical or early symptomatic stages of AD. In addition, given the neurophysiological changes occurring in the pre-clinical stage of the disease, it was expected that participants with positive biomarkers for AD and a genetic risk factor for sporadic AD would exhibit worse performance than healthy controls (Allison et al., 2016; Coughlan et al., 2018).

Thirty-two studies (86.5%) reporting evidence on comparing subgroups of 29 instruments were identified. Seventeen analyses were deemed as very good in quality (Allison et al., 2016; Bierbrauer et al., 2020; Castegnaro et al., 2022; Colmant et al., 2023; Coughlan et al., 2019; Gellersen et al., 2021; Kunz et al., 2015; Laczó et al., 2021, 2022; Morganti et al., 2013; Parizkova et al., 2018; Plaza-Rosales et al., 2015, 2023; Puthusserypady et al., 2022; Serino et al., 2015; Tarnanas et al., 2012, 2015), nine as adequate (Bellassen et al., 2012; Caffò et al., 2012, 2018; Da Costa et al., 2022; Konishi et al., 2018; Migo et al., 2016; Mohammadi et al., 2018; Morganti et al., 2013; Park, 2022), and eight as doubtful (Bayahya et al., 2021; Davis & Sikorskii, 2020; Lesk et al., 2014; Moussavi et al., 2022; Pink et al., 2023; Ritchie et al., 2018; Serino et al., 2018; Zen et al., 2013). The mean size of the groups compared was 34 subjects, while the mean age was 69.24 (± 12.7).

Evidence about the difference in the performance between high and low genetic risk for sporadic AD (i.e., APOE $\epsilon 3\epsilon 3$,

APOE $\epsilon 3\epsilon 4$) was reported in six studies (Bierbrauer et al., 2020; Colmant et al., 2023; Coughlan et al., 2019, 2020; Gellersen et al., 2021; Kunz et al., 2015) and against healthy controls in two studies (Konishi et al., 2018; Pink et al., 2023). Differences in the navigational performance were reported in the four studies, even in the absence of differences in object location memory (Bierbrauer et al., 2020; Colmant et al., 2023; Kunz et al., 2015), memory binding (Gellersen et al., 2021), and path integration tasks (Castegnaro et al., 2022; Coughlan et al., 2019, 2020; Kunz et al., 2015). Spatial navigation alterations seem to be related to the wayfinding ability (Coughlan et al., 2019, 2020; Gellersen et al., 2021; Kunz et al., 2015). This difference in the navigational performance was also reported when the performance between APOE $\epsilon 3\epsilon 4$ carriers and healthy controls was compared in the *concurrent spatial discrimination task* (Konishi et al., 2018) and the *apple game* (Colmant et al., 2023).

The performance of healthy controls was compared with the AD and MCI population in 14 studies (Bellassen et al., 2012; Davis & Sikorskii, 2020; Laczó et al., 2021, 2022; Lee et al., 2014; Mohammadi et al., 2018; Morganti et al., 2013; Parizkova et al., 2018; Puthusserypady et al., 2022; Serino et al., 2015, 2018; Silva et al., 2023; Tarnanas et al., 2015; Zen et al., 2013). As expected, healthy controls showed a better performance in tasks involving planning a path in the presence of allocentric and egocentric perspectives, such as the *VR-maze spatial task* and *virtual reality navigational experiment* (Morganti et al., 2013). Impairment in allocentric and egocentric spatial strategies has been reported in early-stage symptomatic AD and early dementia due to AD samples in multiple virtual reality-based instruments (Morganti et al., 2013; Parizkova et al., 2018; Serino et al., 2015; Zen et al., 2013). Specifically, allocentric impairment has been reported from MCI stages, while egocentric impairments from early AD (Morganti et al., 2013; Parizkova et al., 2018; Ritchie et al., 2018; Ruggiero et al., 2018).

Discussion

Over recent years, technologies such as virtual reality and serious games have progressively expanded in the public health domain (Manera et al., 2017). Using simulated 3D environments to assess spatial navigation and spatial memory is a profitable methodology for accessing clinical samples, such as individuals within the AD continuum (Caffò et al., 2012). This review has identified and evaluated the methodological quality of 30 VRSG-based instruments using the COSMIN framework. Most studies have primarily examined criterion and convergent validity using cross-sectional designs, providing valuable evidence about the diagnostic accuracy and construct validity of the instruments. However, the reliance on such designs, coupled with the use of

non-probabilistic sampling methods, potentially introduces selection and spectrum biases in the results (Knottnerus & Buntinx, 2009; Kohn et al., 2013; Pepe, 2003). This limitation, along with the heterogeneity of theoretical frameworks, complicates the interpretation of findings and hampers the broader clinical application of these instruments. These findings warrant a detailed discussion of the results and implications for future research.

Our findings are in agreement with previous systematic reviews, supporting that VRSG-based instruments demonstrate satisfactory diagnostic accuracy in distinguishing individuals with MCI and AD dementia from those with normal aging (Chan et al., 2021; Molina da Costa et al., 2020; Tuena et al., 2021). While the identified VRSG-based instruments demonstrate strengths in terms of construct measurement and diagnostic accuracy, some questions regarding their content validity and research design have raised broader questions about the theoretical framework and the standardization of development procedures. In the following sections, some of these aspects will be discussed in detail, aiming to identify strategies and future directions that can enhance the existing instruments and facilitate the development of new, more reliable, and valid VRSG-based instruments.

Construct Validity of VRSG-Based Instruments

Construct validity refers to the degree to which the scores of an instrument are consistent with hypotheses, assuming that the instrument validly measures the intended construct (Mokkink et al., 2019). The evaluation of construct validity typically can be investigated using two key procedures: convergent validity and known-group validity (Mokkink et al., 2019). Convergent validity refers to determine the relations with other measures of good quality that are intended to assess the same construct (Mokkink et al., 2019). On the other hand, known-group validity refers to testing the instrument ability to distinguish between groups that are expected to differ based on the construct being measured (Mokkink et al., 2018). Both forms of validity require a precise and clear definition of the construct for formulating accurate hypotheses and selecting appropriate instruments for evaluation (Abma et al., 2016).

Multiple definitions supporting the development of VRSG-based instruments were identified. Constructs such as spatial navigation (Allison et al., 2016; Gellersen et al., 2021; M. Laczó et al., 2022; Levine et al., 2020; Migo et al., 2016; Puthusseryppady et al., 2022; Serino et al., 2015; Tarnanas et al., 2015), spatial memory (Coughlan et al., 2020; Konishi et al., 2018; Park, 2022), spatial cognition (Tarnanas et al., 2012), spatial reorientation (Caffò et al., 2012, 2018), topographic memory (Lesk et al., 2014), object location memory (Kunz et al., 2015), and allocentric and egocentric spatial strategies (Allison

et al., 2019; Caffò et al., 2018; Mohammadi et al., 2018; Morganti et al., 2013) were reported as the conceptual basis supporting the VRSG-based instruments.

Despite the heterogeneity in definitions underlying VRSG-based instruments, most of them can be traced back to the cognitive map hypothesis (Epstein et al., 2017; O'Keefe & Nadel, 1978; Tolman, 1948), according to which the brain builds unified representations of the environment using spatial cues integrated within allocentric and egocentric frameworks (Epstein et al., 2017). Cognitive maps support navigational behavior, including path integration and wayfinding tasks, such as those included in *Sea Hero Quest* (Coughlan et al., 2019; Gellersen et al., 2021; Puthusseryppady et al., 2022), *object location task* (Castegnaro et al., 2022), *VR maze spatial task* (Morganti et al., 2013), *virtual neighborhood task* (Morganti et al., 2013), *virtual supermarket test* (Coughlan et al., 2020; Puthusseryppady et al., 2022), *cognitive mapping task* (Allison et al., 2016; Levine et al., 2020), *route learning task* (Levine et al., 2020), and the *navigation test suite* (Laczó et al., 2021, 2022).

The cognitive map hypothesis provides a comprehensive framework for comprehending the alterations in spatial navigation behavior across the AD continuum, as these changes are closely linked to the neurodegenerative progression of the condition (Coughlan et al., 2018). Extensive research in rodents and humans supports that specific groups of cells in the medial temporal lobe, such as the place, grid, boundary, and head direction cells, play a crucial role in forming these maps (Epstein et al., 2017).

This notion is highlighted in four studies (Bierbrauer et al., 2020; Kunz et al., 2015; Laczó et al., 2022; Parizkova et al., 2018) that show evidence regarding the relationship between the *apple game*, *object location memory task*, *navigation test suite*, *yVSA*, and neuronal activity. The structures included embrace grid cells and volumetry of areas such as the hippocampus and the entorhinal cortex, which are critical for forming cognitive maps (Coughlan et al., 2018; Epstein et al., 2017). This relationship supports that the cognitive map hypothesis is helpful in preclinical AD assessment, based on the evidence that structural changes in crucial areas in the hippocampal formation during the first stages of AD pathology could affect spatial navigation strategies. Specifically, allocentric impairment (Laczó et al., 2018; Nedelska et al., 2012) could be the first cognitive change caused by AD pathology, driven by tau-related degeneration in the medial temporal lobe rather than amyloid deposition in other brain areas such as the medial temporal lobe (Ritchie et al., 2018). Therefore, maintaining the distinction between allocentric and egocentric frameworks while also considering the neuronal progression of the disease could help to develop new cognitive tests that are highly sensitive to the presence of tauopathy in the medial temporal lobe, recognized as a distinctive characteristic of AD (Jack et al., 2018).

The relationship between VRSG-based instruments and neuropsychological tests encompasses various measures of visuospatial processing, such as the *Corsi Test* (Lee et al., 2014; Morganti et al., 2013), the *Complex Figure of Rey-Osterrieth* (Lee et al., 2014), the *Manikin test* (Morganti et al., 2013), the *Money Road-Map test* (Da Costa et al., 2022), and *Benton's Judgment of Line Orientation test* (Da Costa et al., 2022). Additionally, measures of general cognitive status (Da Costa et al., 2022; Lesk et al., 2014; Morganti et al., 2013; Ritchie et al., 2018) have been reported to correlate with VRSG-based instruments. These correlations suggest that VRSG-based instruments are likely related to the spatial cognition domain, involving different spatial abilities such as egocentric mental rotation of space, spatial orientation and transformation, and spatial memory (Carter & Woldstad, 1985; Kessels et al., 2008, 2010; Vingerhoets et al., 1996). Furthermore, the performance on VRSG-based instruments may indicate cognitive detriment as AD progress. To enhance the interpretation of findings in terms of construct validity analyses, future research should report the purpose of correlation analysis.

Regarding the comparison between well-known groups, participants in the MCI stage demonstrated differences in performance compared to healthy controls, affirming the sensitivity of the intended construct in capturing early cognitive impairments. Additionally, including participants with genetic risk for AD further supported the discriminative capacity of the measures (Gellersen et al., 2021). Notably, healthy controls consistently outperformed impaired individuals and those at genetic risk, indicating the progressive nature of spatial navigation and spatial memory impairments along the AD continuum (Castegnaro et al., 2022; Da Costa et al., 2022; Laczó et al., 2021, 2022; Parizkova et al., 2018). However, differences in the classification criteria and metrics used among the studies impose some limitations in comparing results across the literature. The implications derived from the variation in criteria selection are discussed in the subsequent section.

Criterion Validity of VRSG-Based Instruments

Criterion validity is the degree to which the score of an instrument adequately reflects a standard of reference (Mokkink et al., 2018). Therefore, the standard of reference selection should be carefully considered when interpreting its further use in the clinical setting (Liu et al., 2015; Pepe, 2003). The different standards of reference selected in the VRSG-based instruments that evaluated criterion validity limit the comparison of their results. For instance, while *Sea Hero Quest* (Coughlan et al., 2019) showed to be helpful in discriminating between subjects with and without a genetic risk for sporadic AD (i.e., APOE $\epsilon 3\epsilon 4$ carriers vs non-carriers), the *object location* task (Castegnaro et al., 2022), *SOIVET* (Da Costa et al., 2022), *navigation test suite*

(Laczó et al., 2022), *SCT-VR* (Park, 2022), and the *VReoT* (Caffò et al., 2012) showed an adequate capacity to discriminate between MCI and healthy participants.

In the context of AD research, different sources of evidence are needed to determine the presence of the disease and its stage (Jack et al., 2018). According to the National Institute on Aging and Alzheimer's Association research framework, both A β and pathologic tau biomarkers are required for the neuropathologic diagnosis of AD (Jack et al., 2018). In addition, cognitive symptoms are used only to stage the severity of the disease and should not be used to define the presence of AD (Jack et al., 2018).

Five studies presented evidence regarding the criterion validity using biomarker status as a reference (Allison et al., 2016, 2019; Castegnaro et al., 2022; Laczó et al., 2022; Levine et al., 2020). However, four collected A β and tau biomarkers (Allison et al., 2019; Castegnaro et al., 2022; Laczó et al., 2022; Levine et al., 2020) and these were used to define the standard of reference categories in two (Castegnaro et al., 2022; Laczó et al., 2021). In the remaining studies, criterion validity was tested using different categories derived from biomarkers such as high/low level of A β biomarker or clinical assessment (Allison et al., 2016, 2019; Levine et al., 2020). Therefore, only the *object location memory* task (Castegnaro et al., 2022) and the *navigation test suite* (Laczó et al., 2022) have shown validity to distinguish between MCI due to AD and healthy controls based on biomarkers. Further research is needed to analyze the criterion validity of VRSG-based instruments to discriminate between preclinical samples that are cognitively unimpaired according to currently available neuropsychological tests and those with positive biomarkers and healthy controls.

Content Validity and Research Design Requirements

Content validity, which encompasses face validity, is the degree to which the instrument adequately accurately represents the intended construct it aims to measure, as determined by experts' or potential users' assessment (Terwee et al., 2018). Usually, the agreement among experts or pilot studies is accepted as evidence of this psychometric property (Mokkink et al., 2019).

In the absence of agreement among experts, ten of the identified studies reported pilot or previous studies (Allison et al., 2019; Caffò et al., 2012; Davis & Sikorskii, 2020; Laczó et al., 2022; Levine et al., 2020; Parizkova et al., 2018; Puthusserypady et al., 2022; Ritchie et al., 2018; Tarnanas et al., 2012; Zen et al., 2013). Pilot studies provide valuable evidence for designing or refining characteristics of the instruments, such as the number of trials, learning or fatigue effects, and the total duration of the task, as shown by the pilot study of the *virtual reality navigational experiment* (Zen et al., 2013) and the *modified cognitive mapping task* (Allison et al., 2016).

While the COSMIN checklist lacks standardized criteria for assessing usability, it is pertinent to discuss this aspect due to the digital nature of the VRSG-based instruments. Usability, referring to the ease of use and effectiveness of a digital test or tool (Latendorf et al., 2021), is crucial for face validity, particularly in the context of digital interfaces. Using peripherals such as keyboards, joysticks, or gamepads adds an extra layer of difficulty in interacting with the instruments. This interaction should be as clear and easy as possible to prevent any affectation to the performance induced by an unfamiliar usage of the interface or peripheral, as suggested by the literature on recommendations for developing serious games and virtual reality tools for neurodegenerative diseases (Ben-sadoun et al., 2018; Manera et al., 2017). Most of the identified studies reported having included training or practice trials, with the exception of four studies which did not report this information (Davis & Sikorskii, 2020; Gellersen et al., 2021; Kunz et al., 2015; Zen et al., 2013).

In most of the studies reviewed, practice trials, following verbal or written instructions, were employed to familiarize participants with the requirements of the tasks and the use of peripherals. This strategy aimed to ensure a minimum proficiency level, mitigating performance disparities due to difficulties with device handling or task comprehension. While some studies provided usability metrics, such as the number and duration of practice trials, these metrics were not uniformly reported. Furthermore, the performance quality in practice sessions was used as a feasibility criterion, excluding participants who still needed to meet a baseline level of understanding or performance (Castegnaro et al., 2022; Morganti et al., 2013).

Although these measures provide insights into data quality and participant performance, there is a pressing need for further research to assess task and tool usability across the AD continuum. Such studies, particularly at the population level, remain uncommon. Notably, only Silva et al. (2023) presented evidence on the applicability of the *SOIVET* for participants with MCI due to AD, reporting no significant differences in cybersickness symptoms or immersion levels between MCI participants and healthy controls. To comprehensively measure usability, future studies should also consider user experience aspects like perceived usability (Lewis, 2018) and digital ergonomics (Ben-sadoun et al., 2018), incorporating methodologies tailored for user experience assessment (Sauro & Lewis, 2012).

In future studies, spectrum bias in the usability of the instrument should also be considered if pilot studies are carried out using cognitively unimpaired samples. This bias arises as the healthy population may differ in age, educational level, or technology literacy from the preclinical or at-risk population. Therefore, to mitigate potential spectrum bias in usability, especially when pilot studies use

cognitively unimpaired samples, it is crucial to ensure that the target population perceives the instrument as usable. For this, we recommend considering the inclusion of usability questionnaires or focus group methodology and structured interviews in the pilot study (Ben-sadoun et al., 2018; Sauro & Lewis, 2012). The inclusion of such methodologies is recommended to comprehensively assess the usability of VRSG-based instruments, thereby enhancing their applicability in broader settings.

Implications for Clinical Practice

The development of medical tests for screening and diagnosis typically follows a multiphase approach (Knottnerus & Buntinx, 2009; Pepe, 2003; Sackett & Haynes, 2002). The first phases in diagnostic research could be described as exploratory investigations typically following a case-control design, sometimes described as case-referent, and using non-probabilistic samples (Knottnerus & Buntinx, 2009; Pepe, 2003). Usually, these studies are described as phase one or phase two diagnostic studies and try to answer whether the test results in known affected patients differ from those in normal individuals (Sackett & Haynes, 2002). Despite their exploratory nature, this kind of diagnostic research studies provide valuable information, including sensitivity, specificity, likelihood ratios, discriminative capacity (Knottnerus & Buntinx, 2009; Sackett & Haynes, 2002), and usability (Ben-sadoun et al., 2018).

Our findings indicate that VRSG-based instruments align better with early stages of diagnostic research, specifically phases one and two (Sackett & Haynes, 2002). Consequently, their application in clinical screening of cognitive markers related to AD in their current state is not recommended. The limited evidence for diagnostic utility at this stage is primarily due to methodological constraints, such as non-representative sampling and variable reference standards (Pepe, 2003). To establish their efficacy in screening cognitive impairments associated with AD, further research must focus on whether performance in spatial navigation and memory tasks distinctly distinguishes between diverse groups with varying levels of risk for developing AD (Pepe, 2003). Additionally, generating normative data that accounts for variables such as age, education level, and sociodemographic variables is essential for providing an empirical context for interpreting individual performance on VRSG-based tasks (delCacho-Tena et al., 2023). The availability of normative data for VRSG-based instruments could support clinical integration and standardized use in population-level screening programs.

Longitudinal studies are imperative to track spatial navigation in at-risk or diagnosed individuals and should aim to identify dependable diagnostic markers and to evaluate the long-term health outcomes of VRSG-based assessments.

Additionally, research into the cost-effectiveness of integrating these technologies into clinical settings is imperative to assess their feasibility and potential for widespread application. These directions establish a route for the development of fully usable VRSG-based instruments for the clinical screening of cognitive markers related to AD in clinical settings according to diagnostic research recommendations (Knottnerus & Buntinx, 2009; Pepe, 2003).

Despite these challenges, VRSG-based instruments represent the best opportunity for developing an integrated framework for neuropsychological assessment of spatial navigation and spatial memory in humans (Colombo et al., 2017; Coughlan et al., 2019; Park, 2022; Zucchella et al., 2014). Some identified VRSG-based instruments provide a platform for developing cognitive markers usable and translatable to clinical practice in the upcoming years. Instruments such as *Sea Hero Quest* (Coughlan et al., 2019; Puthusseryppady et al., 2022), *navigation test suite* (Laczó et al., 2021, 2022), *virtual supermarket test* (Coughlan et al., 2020; Puthusseryppady et al., 2022), and *object location task* (Castegnaro et al., 2022) provide starting points to build unified paradigms and tools for the development of scalable diagnostic tools and should be considered to new analyses of evidence in future reviews.

Implications for Research

The critical analysis of the existing evidence available sheds light on the future directions for designing and enhancing VRSG-based instruments in the context of early AD detection. Further research is required to establish the psychometric performance and clinical utility of VRSG-based instruments, particularly the content validity, usability, and diagnostic accuracy for preclinical AD. Some of the identified VRSG-based instruments using path integration and wayfinding tasks have demonstrated adequate diagnostic accuracy to discriminate healthy controls, early-stage symptomatic AD, and early dementia due to AD (Allison et al., 2016, 2019; Castegnaro et al., 2022; Laczó et al., 2022; Levine et al., 2020). New research should be focused on testing the criterion validity to discriminate between subjects in preclinical stages of AD and MCI. To enhance the comparability of results, we encourage using the National Institute on Aging and Alzheimer's Association research framework (Jack et al., 2018) as the standard of reference for preclinical AD identification. Other scenarios can include testing VRSG-based instruments in populations with genetic risk factors for sporadic and familial AD (Fuller et al., 2019; Sepulveda-Falla et al., 2012). The innovative approach of utilizing a widely accessible video game, such as *Sea Hero Quest* (Coughlan et al., 2018; Coutrot et al., 2018), has the potential to facilitate the enrollment of participants.

Regarding the theoretical framework supporting VRSG-based instruments, our findings establish the cognitive map

hypothesis (O'Keefe & Nadel, 1978; Tolman, 1948) as the foundational theory in designing VRSG-based instruments and virtual reality paradigms for AD research. This hypothesis is supported by empirical evidence, such as the observed hippocampal neural response during virtual navigation (Epstein et al., 2017). Additionally, other variables related to spatial cognition and general cognitive status could be considered as proxy variables for determining cognitive domain and clinical utility, respectively (Burgess et al., 2004; Hirtle, 2013; Vasilyeva, 2005).

Identified VRSG-based instruments have limitations in providing evidence about the predictive value of the diagnosis and the impact of diagnostic testing on the progression and management of the disease (Knottnerus & Buntinx, 2009; Pepe, 2003). Therefore, new research, including cohort studies, should be considered to determine the operative characteristics of the instruments. Notably, the reliance on common statistics like Cronbach's alpha for reliability assessment is increasingly viewed as insufficient (Sijtsma, 2009). In the context of AD, a degenerative condition with varying cognitive decline time rates (Morrison et al., 2015; Potashman et al., 2023), longitudinal studies are particularly valuable. They can provide deeper insights into the test-retest reliability and responsiveness of VRSG-based instruments (Coughlan et al., 2020; Mokkink et al., 2018), elements crucial for tracking progression of the disease. Given the expected cognitive changes, it is essential to provide a rationale for selecting the recall period to examine potential learning effects and clinical changes associated with disease progression. Following the recommendations for study design suggested by COSMIN (Mokkink et al., 2019) and the considerations for diagnostic research (Knottnerus & Buntinx, 2009; Pepe, 2003) can be helpful to enhance the comparability of results.

In summary, this review underscores the potential of VRSG-based instruments for early AD detection, while also identifying gaps in evidence and critical areas for future research. These instruments, in their current state, stand out as a promising source of cognitive biomarkers for detecting AD from its preclinical stage. However, heterogeneity in research designs and the risk of bias across studies limit their clinical application. Overcoming these challenges calls for joint efforts from clinicians and researchers, advancing these technologies as practical diagnostic tools on a larger scale. For researchers developing VRSG-based instruments, we advise integrating a theoretical framework anchored in neuroscience. The cognitive map hypothesis is particularly suitable, due to its neurological underpinnings and its degree of support within the current knowledge base. Moreover, we strongly recommend that researchers also clearly articulate and rationalize specific hypotheses for construct validity testing, adhere to the NIAA's research framework for AD diagnosis as the most reliable standard, and assess usability through data collection on digital ergonomics and

user interface via qualitative methods, such as interviews, focus groups, or questionnaires (Ben-sadoun et al., 2018; Lewis, 2018; Sauro & Lewis, 2012). Crucially, publishing evidence on validity and usability from early development stages, including pilot studies involving individuals across the AD continuum, will significantly contribute to instrument refinement and effectiveness.

Limitations

To the best of our knowledge, this study represents a pioneering effort in using standardized criteria to analyze the quality of studies on measurement properties for VRSG-based instruments. Nonetheless, some limitations should be recognized. Our adherence to the COSMIN guidelines, which are tailored for PROMs, primarily resulted in a focused yet somewhat restricted scope of evaluation. Notably, the COSMIN framework does not explicitly address aspects such as instrument usability and norming, which are critical for analyzing VRSG-based instruments. Additionally, while ten studies reported pilot studies (Allison et al., 2019; Caffò et al., 2012; Davis & Sikorskii, 2020; Laczó et al., 2022; Levine et al., 2020; Parizkova et al., 2018; Puthusseryppady et al., 2022; Ritchie et al., 2018; Tarnanas et al., 2012; Zen et al., 2013) and one study reported the assessment of applicability (Silva et al., 2023), the lack of standardized criteria for their evaluation limited our analysis. This constraint impeded a thorough assessment of factors influencing instrument usage, such as age and technological literacy. Consequently, our study could not provide an exhaustive set of recommendations for best practices in evaluating these essential aspects of VRSG-based instruments, underlining the need for further research and development of comprehensive assessment criteria in this field.

Our examination focused on studies published from 2012 onwards, across eight major databases. Consequently, investigations reporting on the psychometric properties of VRSG-based instruments published before 2012 are beyond the scope of this work. Although a gray literature search was conducted, it is possible that other sources of evidence still need to be explored. For instance, personal notes and non-published works may contain valuable insights regarding developments, potential paradigms, or pilot studies that could contribute to advancing and refining the VRSG-based instruments. We encourage researchers to publish pilot studies, proof-of-concept, and early designs to advance the standardization and development of virtual reality as feasible technology for human spatial navigation assessment.

The inclusion criteria for this review, which primarily targeted references in English and Spanish, introduce a potential language bias (Pieper & Puljak, 2021). Although English is the predominant language in scientific research

(Meneghini & Packer, 2007) and databases such as SciELO, RedALyC, and LILACS provide access to literature in various languages, limiting our research to only English or Spanish may have inadvertently excluded other contributions. This is particularly true for non-peer-reviewed works published in other languages. This limitation underscores the need for more inclusive linguistic approaches in future research, particularly concerning VRSG-based instruments. Adopting such approaches would ensure a more comprehensive and globally diverse understanding of the subject.

Finally, it is important to note the limitations arising from the heterogeneous selection of reference standards in the included studies. This diversity in diagnostic criteria can affect the comparability of diagnostic accuracy and known-group validity, as different tools may yield varying results depending on the applied standards (Bossuyt et al., 2003). Recognizing this limitation underscores the need for standardized diagnostic criteria in future research, particularly when assessing the sensitivity of spatial navigation paradigms in Alzheimer's disease and MCI.

Conclusion

This systematic review has identified and evaluated 30 VRSG-based instruments, underscoring their potential for assessing spatial navigation and memory impairments from the earliest stages of AD. These instruments are of great clinical importance given that such deficits are among the initial signs of cognitive decline in at-risk populations (Castegnaro et al., 2022; Colombo et al., 2017; Nedelska et al., 2012). While these tools show promise in differentiating between symptomatic stages of AD—including MCI and dementia—from clinically normal individuals, their efficacy in preclinical stages remains to be fully established. Expansion of studies to include familial early-onset AD populations could provide crucial insights into AD pathology, independent of age-related factors, thereby enhancing population-level screening programs.

Our evaluation, as summarized in Table 3, indicates gaps in knowledge and potential biases in existing VRSG-based instruments. These instruments have generally demonstrated construct validity, comparing well-known groups in AD and correlating the performance in spatial navigation and spatial memory tasks with established visuospatial constructs and neural correlates. However, improvements in sampling strategies and statistical hypothesis formulation are needed to strengthen methodological robustness. Criterion validity is particularly challenged by the heterogeneity in standard reference selection, which can be addressed by adhering to contemporary AD research frameworks like that proposed by NIA-AA (Jack et al., 2018) or by providing comprehensive rationales for selecting at-risk groups. Furthermore, the adoption of unified theoretical frameworks will

benefit the development of measurement models aligned with spatial navigation theories, thus enhancing structural validity assessments and contributing to the unification of spatial navigation models.

Longitudinal study designs are essential for establishing the reliability of the measures provided by VRSG-based instruments over time. The selection of follow-up times must be carefully rationalized to ascertain the reliability of an instrument. Improving the reporting of methodological details in studies is crucial for reducing risk of bias and enhancing the understanding of study designs (Mokkink et al., 2019). Therefore, researchers should adhere to guidelines like STARD for reporting diagnostic accuracy in dementia research (Noel-Storr et al., 2014), ensuring clarity in hypotheses, appropriate sampling methods, and adequate sample sizes for statistical power.

In conclusion, effectively addressing these identified gaps will significantly improve decision-making processes regarding the adoption of VRSG-based instruments as diagnostic tools for AD. By adhering to the study recommendations (outlined in supplementary material), researchers have an opportunity to substantially contribute to the evolution and refinement of VRSG-based instruments, advancing diagnostics in both the clinical and research domains of AD.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11065-024-09633-7>.

Author Contribution JPS conceived the review idea and conceptualized the search. MGB, DCA, and NT contributed equally to the study formulation and the conceptualization of the protocol. DCA provided expertise in the methodological and technical aspects of the search strategy and the synthesis of information. MGB provided guidance on the qualitative synthesis of information, particularly in the context of measurement properties of neuropsychological instruments. NT made significant contributions to the research development, offering close mentorship in manuscript writing and critical analysis of the findings. AMG, DST, LCT, CJK, and JPS searched, synthesized the findings, and analyzed the risk of bias in the identified studies. JPS wrote the first draft of the manuscript, and all authors commented on previous versions. All authors contributed substantially to the revision of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding provided by Colombia Consortium Ministerio de Ciencia, Tecnología e Innovación, 111591891726. Contract: 588-2022.

Data Availability The datasets and tables generated during the review are available in Open Science Framework repository, https://osf.io/fyx4/?view_only=34c89a393d294ec29298c836b8c4aee4.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abma, I. L., Rovers, M., & Van Der Wees, P. J. (2016). Appraising convergent validity of patient-reported outcome measures in systematic reviews: Constructing hypotheses and interpreting outcomes. *BMC Research Notes*, 9(226), 1–5. <https://doi.org/10.1186/s13104-016-2034-2>
- Allison, S. L., Fagan, A. M., Morris, J. C., & Head, D. (2016). Spatial navigation in preclinical Alzheimer's disease Samantha. *Journal of Alzheimer's Disease*, 52(1), 77–90. <https://doi.org/10.3233/JAD-150855>
- Allison, S. L., Rodebaugh, T. L., Johnston, C., Fagan, A. M., Morris, J. C., & Head, D. (2019). Developing a spatial navigation screening tool sensitive to the preclinical Alzheimer disease continuum. *Archives of Clinical Neuropsychology*, 34(7), 1138–1155. <https://doi.org/10.1093/arclin/acz019>
- Alsbury-Nealy, K., Wang, H., Howarth, C., Gordienko, A., Schlichting, M. L., & Duncan, K. D. (2021). OpenMaze: An open-source toolbox for creating virtual navigation experiments. *Behavior Research Methods*, 54(3), 1374–1387. <https://doi.org/10.3758/s13428-021-01664-9>
- Alzheimer's Association. (2023). *2023 Alzheimer's disease facts and figures*, 19(4), 1598–1695. <https://doi.org/10.1002/alz.13016>
- Bayahya, A. Y., Alhalabi, W., & Alamri, S. H. (2021). Smart health system to detect dementia disorders using virtual reality. *Healthcare (Switzerland)*, 9(7), 810. <https://doi.org/10.3390/healthcare9070810>
- Bellassen, V., Iglói, K., de Souza, L. C., Dubois, B., & Rondi-Reig, L. (2012). Temporal order memory assessed during spatiotemporal navigation as a behavioral cognitive marker for differential Alzheimer's disease diagnosis. *Journal of Neuroscience*, 32(6), 1942–1952. <https://doi.org/10.1523/JNEUROSCI.4556-11.2012>
- Ben-sadoun, G., Manera, V., Alvarez, J., Sacco, G., Robert, P., Kuljiš, R. O., & García-betances, R. I. (2018). Recommendations for the design of serious games in neurodegenerative diseases. *Frontiers in Aging Neuroscience*, 10(February), 1–7. <https://doi.org/10.3389/fnagi.2018.00013>
- Bierbrauer, A., Kunz, L., Gomes, C. A., Luhmann, M., Deuker, L., Getzmann, S., Wascher, E., Gajewski, P. D., Hengstler, J. G., Fernandez-Alvarez, M., Atienza, M., Cammisuli, D. M., Bonatti, F., Pruneti, C., Percesepe, A., Bellaali, Y., Hanseeuw, B., Strange, B. A., Cantero, J. L., & Axmacher, N. (2020). Unmasking selective path integration deficits in Alzheimer's disease risk carriers. *Science Advances*, 6(35). <https://doi.org/10.1126/sciadv.aba1394>
- Bird, C. M., & Burgess, N. (2009). Spatial memory: Assessment in animals. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience* (pp.

- 187–194). Elsevier. <https://doi.org/10.1016/B978-008045046-9.00288-6>
- Boot, W. R. (2015). Video games as tools to achieve insight into cognitive processes. *Frontiers in Psychology*, 6(1), 1–3. <https://doi.org/10.3389/fpsyg.2015.00003>
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Moher, D., Rennie, D., de Vet, H. C., Lijmer, J. G., & Standards for Reporting of Diagnostic Accuracy. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Annals of Internal Medicine*, 138(1), W1–W12. <https://doi.org/10.7326/0003-4819-138-1-200301070-00012-w1>
- Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12), 551–557. <https://doi.org/10.1016/j.tics.2006.10.005>
- Burgess, N., Spiers, H. J., & Paleologou, E. (2004). Orientational manoeuvres in the dark: Dissociating allocentric and egocentric influences on spatial memory. *Cognition*, 94(2), 149–166. <https://doi.org/10.1016/j.cognition.2004.01.001>
- Caffò, A. O., De Caro, M. F., Picucci, L., Notarnicola, A., Settanni, A., Livrea, P., Lancioni, G. E., & Bosco, A. (2012). Reorientation deficits are associated with amnesic mild cognitive impairment. *American Journal of Alzheimer's Disease and Other Dementias*, 27(5), 321–330. <https://doi.org/10.1177/1533317512452035>
- Caffò, A. O., Lopez, A., Spano, G., Serino, S., Cipresso, P., Stasolla, F., Savino, M., Lancioni, G. E., Riva, G., & Bosco, A. (2018). Spatial reorientation decline in aging: The combination of geometry and landmarks. *Aging and Mental Health*, 22(10), 1372–1383. <https://doi.org/10.1080/13607863.2017.1354973>
- Campbell, Z., Zakzanis, K. K., Jovanovski, D., Joordens, S., Mraz, R., & Graham, S. J. (2009). Utilizing virtual reality to improve the ecological validity of clinical neuropsychology: An fMRI case study elucidating the neural basis of planning by comparing the Tower of London with a three-dimensional navigation task. *Applied Neuropsychology*, 16(4), 295–306. <https://doi.org/10.1080/09084280903297891>
- Carter, R., & Woldstad, J. (1985). Repeated measurements of spatial ability with the Manikin test. *HUMAN FACTORS*, 27(2), 209–219. <https://doi.org/10.1177/001872088502700208>
- Castegnaro, A., Howett, D., Li, A., Harding, E., Chan, D., Burgess, N., & King, J. (2022). Assessing mild cognitive impairment using object-location memory in immersive virtual environments. *Hippocampus*, 32(9), 660–678. <https://doi.org/10.1002/hipo.23458>
- Chan, J. Y. C., Yau, S. T. Y., Kwok, T. C. Y., & Tsoi, K. K. F. (2021). Diagnostic performance of digital cognitive tests for the identification of MCI and dementia: A systematic review. *Ageing Research Reviews*, 72, 101506. <https://doi.org/10.1016/j.arr.2021.101506>
- Cipresso, P., Giglioli, I. A. C., Raya, M. A., & Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2018.02086>
- Cogné, M., Taillade, M., N'Kaoua, B., Tarruella, A., Klinger, E., Larrue, F., Sauzéon, H., Joseph, P. A., & Sorita, E. (2017). The contribution of virtual reality to the diagnosis of spatial navigation disorders and to the study of the role of navigational aids: A systematic literature review. *Annals of Physical and Rehabilitation Medicine*, 60(3), 164–176. <https://doi.org/10.1016/j.rehab.2015.12.004>
- Colmant, L., Bierbrauer, A., Bellaali, Y., Kunz, L., Van Dongen, J., Slegers, K., Axmacher, N., Lefèvre, P., & Hanseeuw, B. (2023). Dissociating effects of aging and genetic risk of sporadic Alzheimer's disease on path integration. *Neurobiology of Aging*, 131, 170–181. <https://doi.org/10.1016/j.neurobiolaging.2023.07.025>
- Colombo, D., Serino, S., Tuena, C., Pedroli, E., Dakanalis, A., Cipresso, P., & Riva, G. (2017). Egocentric and allocentric spatial reference frames in aging: A systematic review. *Neuroscience and Biobehavioral Reviews*, 80, 605–621. <https://doi.org/10.1016/j.neubiorev.2017.07.012>
- Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A. M., Spiers, H., & Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 116(19), 9285–9292. <https://doi.org/10.1073/pnas.1901600116>
- Coughlan, G., Laczó, J., Hort, J., Minihane, A. M., & Hornberger, M. (2018). Spatial navigation deficits — Overlooked cognitive marker for preclinical Alzheimer disease? *Nature Reviews Neurology*, 14(8), 496–506. <https://doi.org/10.1038/s41582-018-0031-x>
- Coughlan, G., Puthusseryppady, V., Lowry, E., Gillings, R., Spiers, H., Minihane, A. M., & Hornberger, M. (2020). Test-retest reliability of spatial navigation in adults at-risk of Alzheimer's disease. *PLoS ONE*, 15(9), e0239077. <https://doi.org/10.1371/journal.pone.0239077>
- Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V. D., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2018). Global determinants of navigation ability. *Current Biology*, 28(17), 2861–2866.e4. <https://doi.org/10.1016/j.cub.2018.06.009>
- Da Costa, R. Q. M., Pompeu, J. E., Moretto, E., Silva, J. M., Dos Santos, M. D., Nitri, R., & Brucki, S. M. D. (2022). Two immersive virtual reality tasks for the assessment of spatial orientation in older adults with and without cognitive impairment: Concurrent validity, group comparison, and accuracy results. *Journal of the International Neuropsychological Society*, 28(5), 460–472. <https://doi.org/10.1017/S1355617721000655>
- Davis, R., & Sikorski, A. (2020). Eye tracking analysis of visual cues during wayfinding in early stage Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 49(1), 91–97. <https://doi.org/10.1159/000506859>
- delCacho-Tena, A., Christ, B. R., Arango-Lasprilla, J. C., Perrin, P. B., Rivera, D., & Olabarrieta-Landa, L. (2023). Normative data estimation in neuropsychological tests: A systematic review. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 00, 1–16. <https://doi.org/10.1093/arclin/acad084>
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining “gamification.” *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek 2011*, 9–15. <https://doi.org/10.1145/2181037.2181040>
- Diersch, N., & Wolbers, T. (2019). The potential of virtual reality for spatial navigation research across the adult lifespan. *Journal of Experimental Biology*, 222(Suppl_1). <https://doi.org/10.1242/jeb.187252>
- Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron*, 83(4), 764–770. <https://doi.org/10.1016/j.neuron.2014.07.032>
- Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20(11), 1504–1513. <https://doi.org/10.1038/nn.4656>
- Fernandez-Baizan, C., Arias, J. L., & Mendez, M. (2020a). Spatial memory assessment reveals age-related differences in egocentric and allocentric memory performance. *Behavioural Brain Research*, 388, 112646. <https://doi.org/10.1016/j.bbr.2020.112646>
- Fernandez-Baizan, C., Diaz-Caceres, E., Arias, J. L., & Mendez, M. (2019). Egocentric and allocentric spatial memory in healthy aging: Performance on real-world tasks. *Brazilian Journal of*

- Medical and Biological Research*, 52(4), 1–7. <https://doi.org/10.1590/1414-431x20198041>
- Fernandez-Baizan, C., Nuñez, P., Arias, J. L., & Mendez, M. (2020b). Egocentric and allocentric spatial memory in typically developed children: Is spatial memory associated with visuospatial skills, behavior, and cortisol? *Brain and Behavior*, 10(5), 1–14. <https://doi.org/10.1002/brb3.1532>
- Fuller, J. T., Cronin-golomb, A., Gatchel, J. R., Norton, D. J., Guzmán-vélez, E., Jacobs, H. I. L., Pardilla-delgado, E., Artola, A., Baena, A., Bocanegra, Y., Kosik, K., Chen, K., Tariot, P. N., Sperling, R. A., Reiman, E. M., Lopera, F., & Quiroz, Y. T. (2019). Biological and cognitive markers of presenilin1 e280a autosomal dominant Alzheimer's disease: A comprehensive review of the Colombian kindred. *The Journal of Prevention of Alzheimer's Disease - JPAD*, 6(2), 112–120. <https://doi.org/10.14283/jpad.2019.6.Biological>
- Gazova, I., Vlcek, K., Laczó, J., Nedelska, Z., Hyncicova, E., Mokrisova, I., Sheardova, K., & Hort, J. (2012). Spatial navigation—A unique window into physiological and pathological aging. *Frontiers in Aging Neuroscience*. <https://doi.org/10.3389/fnagi.2012.00016>
- Gellersen, H. M., Coughlan, G., Hornberger, M., & Simons, J. S. (2021). Memory precision of object-location binding is unimpaired in APOE ε4-carriers with spatial navigation deficits. *Brain Communications*, 3(2). <https://doi.org/10.1093/braincomms/fcab087>
- Goodroe, S. C., Starnes, J., & Brown, T. I. (2018). The complex nature of hippocampal-striatal interactions in spatial navigation. *Frontiers in Human Neuroscience*, 12, 250. <https://doi.org/10.3389/fnhum.2018.00250>
- Hirtle, S. C. (2013). Models of spatial cognition. In D. Waller, & L. Nadel (Eds.), *Handbook of spatial cognition* (pp. 211–226). American Psychological Association.
- Iachini, T., Iavarone, A., Senese, V. P., Ruotolo, F., & Ruggiero, G. (2009). *Visuospatial Memory in Healthy Elderly AD and MCI: A Review*, 2(1), 43–59. <https://doi.org/10.2174/1874609810902010043>
- Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Budd Haeberlein, S., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T. J., Phelps, C. H., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., & Sperling, R. A. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dementia*, 14(4), 535–562. <https://doi.org/10.1016/j.jalz.2018.02.018.NIA-AA>
- Jacobs, L. F. (2003). Memory, spatial. In M. J. Aminoff, & R. B. Daroff (Eds.), *Encyclopedia of the neurological sciences* (pp. 86–90). Academic Press. <https://doi.org/10.1016/B0-12-226870-9/01375-7>
- Kessels, R. P. C., Overbeek, A., & Bouman, Z. (2015). Assessment of verbal and visuospatial working memory in mild cognitive impairment and Alzheimer's dementia. *Dementia & Neuropsychologia*, 9(3), 301–305. <https://doi.org/10.1590/1980-57642015DN93000014>
- Kessels, R. P. C., Van Den Berg, E., Ruis, C., & Brands, A. M. A. (2008). The backward span of the Corsi block-tapping task and its association with the WAIS-III digit span. *Assessment*, 15(4), 426–434. <https://doi.org/10.1177/1073191108315611>
- Kessels, R. P. C., Van Zandvoort, M. J. E., Postma, A., Jaap, L., De Haan, E. H. F., Kessels, R. P. C., Van Zandvoort, M. J. E., Postma, A., Jaap, L., Kessels, R. P. C., & Postma, A. (2010). The Corsi block-tapping task: Standardization and normative data. *Applied Neuropsychology*, 4282(7), 252–258. <https://doi.org/10.1207/S15324826AN0704>
- Knottnerus, J. A., & Buntinx, F. (2009). The evidence base of clinical diagnosis: Theory and methods of diagnostic research: Second edition. In *The evidence base of clinical diagnosis: Theory and methods of diagnostic research: Second edition*. <https://doi.org/10.1002/9781444300574>
- Kohn, M. A., Carpenter, C. R., & Newman, T. B. (2013). Understanding the direction of bias in studies of diagnostic test accuracy. *Academic Emergency Medicine*, 20(11), 1194–1206. <https://doi.org/10.1111/acem.12255>
- Kolarik, B., & Ekstrom, A. (2015). The neural underpinnings of spatial memory and navigation. In A. W. Toga (Ed.), *Brain Mapping* (pp. 507–514). Academic Press. <https://doi.org/10.1016/B978-0-12-397025-1.00277-3>
- Konishi, K., Joobor, R., Poirier, J., MacDonald, K., Chakravarty, M., Patel, R., Breitner, J., & Bohbot, V. D. (2018). Healthy versus entorhinal cortical atrophy identification in asymptomatic APOE4 carriers at risk for Alzheimer's disease. *Journal of Alzheimer's Disease*, 61(4), 1493–1507. <https://doi.org/10.3233/JAD-170540>
- Krohn, S., Tromp, J., Quinque, E. M., Belger, J., Klotzsche, F., Rekers, S., Chojacki, P., de Mooij, J., Akbal, M., McCall, C., Villringer, A., Gaebler, M., Finke, C., & Thöne-Otto, A. (2020). Multi-dimensional evaluation of virtual reality paradigms in clinical neuropsychology: Application of the VR-Check framework. *Journal of Medical Internet Research*, 22(4), e16724. <https://doi.org/10.2196/16724>
- Kunz, L., Brandt, A., Reinacher, P. C., Staresina, B. P., Reifensstein, E. T., Weidemann, C. T., Herweg, N. A., Patel, A., Tsitsiklis, M., Kempter, R., Kahana, M. J., Schulze-Bonhage, A., & Jacobs, J. (2021). A neural code for egocentric spatial maps in the human medial temporal lobe. *Neuron*, 109(17), 2781–2796.e10. <https://doi.org/10.1016/j.neuron.2021.06.019>
- Kunz, L., Navarro Schröder, T., Lee, H., Montag, C., Lachmann, B., Sariyska, R., Reuter, M., Stirnberg, R., Stöcker, T., Messing-Floeter, P. C., Fell, J., Doeller, C. F., & Axmacher, N. (2015). Reduced grid-cell-like representations in adults at genetic risk for Alzheimer's disease. *Science*, 350(6259), 430–433. <https://doi.org/10.1126/science.aad1171>
- Laczó, J., Andel, R., Vyhnaek, M., Vlcek, K., Magerova, H., Varjassyova, A., Nedelska, Z., Gazova, I., Bojar, M., Sheardova, K., & Hort, J. (2012). From Morris water maze to computer tests in the prediction of Alzheimer's disease. *Neurodegenerative Diseases*, 10(1–4), 153–157. <https://doi.org/10.1159/000333121>
- Laczó, J., Andel, R., Vyhnaek, M., Vlcek, K., Nedelska, Z., Matoska, V., Gazova, I., Mokrisova, I., Sheardova, K., & Hort, J. (2014). APOE and spatial navigation in Amnesic MCI: Results from a computer-based test. *Neuropsychology*, 28(5), 676–684. <https://doi.org/10.1037/neu0000072>
- Laczó, J., Parizkova, M., & Moffat, S. D. (2018). Spatial navigation, aging and Alzheimer's disease. *Aging*, 10(11), 3050–3051. <https://doi.org/10.18632/aging.101634>
- Laczó, M., Martinkovic, L., Lerch, O., Wiener, J. M., Kalinova, J., Matuskova, V., Nedelska, Z., Vyhnaek, M., Hort, J., & Laczó, J. (2022). Different profiles of spatial navigation deficits in Alzheimer's disease biomarker-positive versus biomarker-negative older adults with amnesic mild cognitive impairment. *Frontiers in Aging Neuroscience*. <https://doi.org/10.3389/fnagi.2022.886778>
- Laczó, M., Wiener, J. M., Kalinova, J., Matuskova, V., Vyhnaek, M., Hort, J., & Laczó, J. (2021). *Spatial Navigation and Visuospatial Strategies in Typical and Atypical Aging*. <https://doi.org/10.3390/brainsci>
- Latendorf, A., Runde, L. M., Salminen, T., & Steinert, A. (2021). Digitization of neuropsychological diagnostics: A pilot study to compare three paper-based and digitized cognitive assessments. *Aging Clinical and Experimental Research*, 33(6), 1585–1597. <https://doi.org/10.1007/s40520-020-01668-z>

- Lee, J., Kho, S., Yoo, H. B., Park, S., & Choi, J. (2014). Spatial memory impairments in amnesic mild cognitive impairment in a virtual radial arm maze. *Neuropsychiatric Disease and Treatment*, 10, 653–660. <https://doi.org/10.2147/NDT.S58185>
- Lesk, V. E., Wan Shamsuddin, S. N., Walters, E. R., & Ugail, H. (2014). Using a virtual environment to assess cognition in the elderly. *Virtual Reality*, 18(4), 271–279. <https://doi.org/10.1007/s10055-014-0252-2>
- Lester, A. W., Moffat, S. D., Wiener, J. M., Barnes, C. A., & Wolbers, T. (2017). The aging navigational system. *Neuron*, 95(5), 1019–1035. <https://doi.org/10.1016/j.neuron.2017.06.037>
- Levine, T. F., Allison, S. L., Stojanovic, M., Fagan, A. M., Morris, J. C., & Head, D. (2020). Spatial navigation ability predicts progression of dementia symptomatology. *Alzheimer's and Dementia*, 16(3), 491–500. <https://doi.org/10.1002/alz.12031>
- Levy, B., Tsoy, E., & Gable, S. (2016). Developing cognitive markers of Alzheimer's disease for primary care: Implications for behavioral and global prevention. *Journal of Alzheimer's Disease*, 54(4), 1259–1272. <https://doi.org/10.3233/JAD-160309>
- Lewis, J. R. (2018). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction*, 34(12), 1148–1156. <https://doi.org/10.1080/10447318.2017.1418805>
- Liu, Y., Chen, Y., & Chu, H. (2015). A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard. *Biometrics*, 71(2), 538–547. <https://doi.org/10.1111/biom.12264>
- Lizio, R. (2020). Classification of Alzheimer's disease dementia (ADD) by using features derived from resting-state electroencephalography (rsEEG). *Alzheimer's & Dementia*, 16(S4). <https://doi.org/10.1002/alz.045816>
- Manera, V., Ben-sadoun, G., Aalbers, T., Agopyan, H., Askenazy, F., Benoit, M., Bensamoun, D., Bourgeois, J., Bredin, J., Bremond, F., Crispim-junior, C., & David, R. (2017). Recommendations for the use of serious games in neurodegenerative disorders: 2016 Delphi panel. *Frontiers in Psychology*, 8(July), 1–10. <https://doi.org/10.3389/fpsyg.2017.01243>
- Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? *EMBO Reports*, 8(2), 112–116. <https://doi.org/10.1038/sj.embor.7400906>
- Migo, E. M., O'Daly, O., Mitterschiffthaler, M., Antonova, E., Dawson, G. R., Dourish, C. T., Craig, K. J., Simmons, A., Wilcock, G. K., McCulloch, E., Jackson, S. H. D., Kopelman, M. D., Williams, S. C. R., & Morris, R. G. (2016). Investigating virtual reality navigation in amnesic mild cognitive impairment using fMRI. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 23(2), 196–217. <https://doi.org/10.1080/13825585.2015.1073218>
- Mitolo, M., Gardini, S., Fasano, F., Crisi, G., & Pelosi, A. (2013). visuospatial memory and neuroimaging correlates in mild cognitive impairment. *Journal of*, 35, 75–90. <https://doi.org/10.3233/JAD-121288>
- Mohammadi, A., Kargar, M., & Hesami, E. (2018). Using virtual reality to distinguish subjects with multiple- but not single-domain amnesic mild cognitive impairment from normal elderly subjects. *Psychogeriatrics*, 18(2), 132–142. <https://doi.org/10.1111/psyg.12301>
- Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Terwee, C. B. (2019). COSMIN study design checklist for patient-reported outcome measurement instruments. *Department of Epidemiology and Biostatistics Amsterdam Public Health Research Institute Amsterdam University Medical Centers, Location VUmc, July*, 1–32. Retrieved October 11, 2022, from www.cosmin.nl
- Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 27(5), 1171–1179. <https://doi.org/10.1007/s11136-017-1765-4>
- Molina da Costa, R. Q., Pompeu, J. E., Pereira de Viveiro, L. A., & Dozzi Brucki, S. M. (2020). Spatial orientation tasks show moderate to high accuracy for the diagnosis of mild cognitive impairment: A systematic literature review. *Arquivos De Neuro-Psiquiatria*, 78, 713–723. <https://doi.org/10.1590/0004-282X20200043>
- Morganti, F., Stefanini, S., & Riva, G. (2013). From allo- to egocentric spatial ability in early Alzheimer's disease: A study with virtual reality spatial tasks. *Cognitive Neuroscience*, 4(3–4), 171–180. <https://doi.org/10.1080/17588928.2013.854762>
- Morrison, G. E., Simone, C. M., Ng, N. F., Hardy, J. L., & Morrison, G. E. (2015). Reliability and validity of the NeuroCognitive Performance Test, a web-based neuropsychological assessment. *Frontiers in psychology*, 6(November), 1–15. <https://doi.org/10.3389/fpsyg.2015.01652>
- Motes, M. A., Finlay, G. A., & Kozhevnikov, M. (2006). Scene recognition following locomotion around a scene. *Perception*, 35(11), 1507–1520. <https://doi.org/10.1068/p5459>
- Moussavi, Z., Kimura, K., & Lithgow, B. (2022). Egocentric spatial orientation differences between Alzheimer's disease at early stages and mild cognitive impairment: A diagnostic aid. *Medical and Biological Engineering and Computing*, 60(2), 501–509. <https://doi.org/10.1007/s11517-021-02478-9>
- Nedelska, Z., Andel, R., Laczó, J., Vlcek, K., Horinek, D., Lisy, J., Sheardova, K., Bures, J., & Hort, J. (2012). Spatial navigation impairment is proportional to right hippocampal volume. *Proceedings of the National Academy of Sciences of the United States of America*, 109(7), 2590–2594. <https://doi.org/10.1073/pnas.1121588109>
- Nguyen, K. V., Tansan, M., & Newcombe, N. S. (2023). Studying the development of navigation using virtual environments. *Journal of Cognition and Development*, 24(1), 1–16. <https://doi.org/10.1080/15248372.2022.2133123>
- Noel-Storr, A. H., McCleery, J. M., Richard, E., Ritchie, C. W., Flicker, L., Cullum, S. J., Davis, D., Quinn, T. J., Hyde, C., Rutjes, A. W. S., Smailagic, N., Marcus, S., Black, S., Blennow, K., Brayne, C., Fiorivanti, M., Johnson, J. K., Köpke, S., Schneider, L. S., ... & McShane, R. (2014). Reporting standards for studies of diagnostic test accuracy in dementia: The STARDdem Initiative. *Neurology*, 83(4), 364–373. <https://doi.org/10.1212/WNL.0000000000000621>
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press.
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 1–10. <https://doi.org/10.1186/s13643-016-0384-4>
- Pan, Z., Cheok, A. D., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. *Computers and Graphics (pergamon)*, 30(1), 20–28. <https://doi.org/10.1016/j.cag.2005.10.004>
- Parizkova, M., Lerch, O., Moffat, S. D., Andel, R., Mazancova, A. F., Nedelska, Z., Vyhnaek, M., Hort, J., & Laczó, J. (2018). The effect of Alzheimer's disease on spatial navigation strategies. *Neurobiology of Aging*, 64, 107–115. <https://doi.org/10.1016/j.neurobiolaging.2017.12.019>
- Park, J. H. (2022). Can the virtual reality-based spatial memory test better discriminate mild cognitive impairment than neuropsychological assessment? *International Journal of Environmental Research and Public Health*, 19(16). <https://doi.org/10.3390/ijerph19169950>
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction (Oxford Sta)*. Oxford University Press.

- Pieper, D., & Puljak, L. (2021). Language restrictions in systematic reviews should not be imposed in the search strategy but in the eligibility criteria if necessary. *Journal of Clinical Epidemiology*, 132, 146–147. <https://doi.org/10.1016/j.jclinepi.2020.12.027>
- Pink, D., Ilkel, E., Chandreswaran, V., Moser, D., Getzmann, S., Patrick, G., Axmacher, N., & Zhang, H. (2023). Modeling the impact of genotype, age, sex, and continuous navigation on pathway integration performance. *BioRxiv*. <https://doi.org/10.1101/2023.09.11.556925>
- Plaza-Rosales, I., Brunetti, E., Montefusco-Siegmund, R., Madariaga, S., Hafelin, R., Ponce, D. P., Behrens, M. I., Maldonado, P. E., & Paula-Lima, A. (2023). Visual-spatial processing impairment in the occipital-frontal connectivity network at early stages of Alzheimer's disease. *Frontiers in Aging Neuroscience*. <https://doi.org/10.3389/fnagi.2023.1097577>
- Poos, J. M., van der Ham, I. J. M., Leeuwis, A. E., Pijnenburg, Y. A. L., van der Flier, W. M., & Postma, A. (2021). Short digital spatial memory test detects impairment in Alzheimer's disease and mild cognitive impairment. *Brain Sciences*, 11(10), 1350. <https://doi.org/10.3390/brainsci11101350>
- Potashman, M., Pang, M., Tahir, M., Shahraz, S., Dichter, S., Pernecky, R., & Nolte, S. (2023). Psychometric properties of the Alzheimer's Disease Cooperative Study – Activities of Daily Living for Mild Cognitive Impairment (ADCS-MCI-ADL) scale: A post hoc analysis of the ADCS ADC-008 trial. *BMC Geriatrics*, 23(1). <https://doi.org/10.1186/s12877-022-03527-0>
- Prinsen, C. A. C., Mookink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 27(5), 1147–1157. <https://doi.org/10.1007/s11136-018-1798-3>
- Puthusserypady, V., Morrissey, S., Spiers, H., Patel, M., & Hornberger, M. (2022). Predicting real world spatial disorientation in Alzheimer's disease patients using virtual reality navigation tests. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-17634-w>
- Ribordy, F., Jabès, A., Banta, P., & Lavenex, P. (2013). Development of allocentric spatial memory abilities in children from 18 months to 5 years of age. *Cognitive Psychology*, 66(1), 1–29. <https://doi.org/10.1016/j.cogpsych.2012.08.001>
- Ritchie, K., Carrière, I., Howett, D., Su, L., Hornberger, M., O'Brien, J. T., Ritchie, C. W., & Chan, D. (2018). Allocentric and egocentric spatial processing in middle-aged adults at high risk of late-onset Alzheimer's disease: The PREVENT dementia study. *Journal of Alzheimer's Disease*, 65(3), 885–896. <https://doi.org/10.3233/JAD-180432>
- Ruggiero, G., Iavarone, A., & Iachini, T. (2018). Allocentric to egocentric spatial switching: Impairment in aMCI and Alzheimer's disease patients? *Current Alzheimer Research*, 15, 229–236. <https://doi.org/10.2174/1567205014666171030114821>
- Sabbagh, M. N., & Blennow, K. (2019). Peripheral biomarkers for Alzheimer's disease: Update and progress. *Neurology and Therapy*, 8(S2), 33–36. <https://doi.org/10.1007/s40120-019-00171-6>
- Sacco, G., Ben-Sadoun, G., Bourgeois, J., Fabre, R., Manera, V., & Robert, P. (2019). Comparison between a paper-pencil version and computerized version for the realization of a neuropsychological test: The example of the trail making test. *Journal of Alzheimer's Disease*, 68(4), 1657–1666. <https://doi.org/10.3233/JAD-180396>
- Sackett, D. L., & Haynes, R. B. (2002). The architecture of diagnostic research. *BMJ*, 324(7336), 539–541. <https://doi.org/10.1136/bmj.324.7336.539>
- Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Elsevier Inc.
- Schöberl, F., Pradhan, C., Irving, S., Buerger, K., Xiong, G., Kugler, G., Kohlbecher, S., Engmann, J., Werner, P., Brendel, M., Schneider, E., Pernecky, R., Jahn, K., la Fougère, C., Bartenstein, P., Brandt, T., Dieterich, M., & Zwergal, A. (2020). Real-space navigation testing differentiates between amyloid-positive and -negative aMCI. *Neurology*, 94(8), e861–e873. <https://doi.org/10.1212/WNL.0000000000008758>
- Sepulveda-Falla, D., Glatzel, M., & Lopera, F. (2012). Phenotypic profile of early-onset familial Alzheimer's disease caused by presenilin-1 E280A mutation. *Journal of Alzheimer's Disease*, 32(1), 1–12. <https://doi.org/10.3233/JAD-2012-120907>
- Serino, S., Morganti, F., Colombo, D., & Riva, G. (2018). The contribution of allocentric impairments to the cognitive decline in Alzheimer's disease. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 253, 84–91. https://doi.org/10.1007/978-3-030-01093-5_11
- Serino, S., Morganti, F., Di Stefano, F., & Riva, G. (2015). Detecting early egocentric and allocentric impairments deficits in Alzheimer's disease: An experimental study with virtual reality. *Frontiers in Aging Neuroscience*, 7, 1–10. <https://doi.org/10.3389/fnagi.2015.00088>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Silva, J. M. D., dos Santos, M. D., da Costa, R. Q. M., Moretto, E. G., de Viveiro, L. A. P., de Lopes, R. D., Brucki, S. M. D., & Pompeu, J. E. (2023). Applicability of an immersive virtual reality system to assess egocentric orientation of older adults. *Arquivos De Neuro-Psiquiatria*, 81(01), 019–026. <https://doi.org/10.1055/s-0042-1759762>
- Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-bogorad, M., ... & Phelps, C. H. (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. *Alzheimer's & Dementia*, 7(3), 280–292. <https://doi.org/10.1016/j.jalz.2011.03.003>
- Sperling, R., Mormino, E., & Johnson, K. (2014). The evolution of preclinical Alzheimer's disease: Implications for prevention trials. *Neuron*, 84(3), 608–622. <https://doi.org/10.1016/j.neuron.2014.10.038>
- Tarnanas, I., Laskaris, N., & Tsolaki, M. (2012). On the comparison of VR-responses, as performance measures in prospective memory, with auditory P300 responses in MCI detection. *Studies in Health Technology and Informatics*, 181, 156–161.
- Tarnanas, I., Papagiannopoulos, S., Kazis, D., Wiederhold, M., Wiederhold, B., Vuillermot, S., & Tsolaki, M. (2015). Reliability of a novel serious game using dual-task gait profiles to early characterize aMCI. *Frontiers in Aging Neuroscience*. <https://doi.org/10.3389/fnagi.2015.00050>
- Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mookink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 27(5), 1159–1170. <https://doi.org/10.1007/s11136-018-1829-0>
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208. <https://doi.org/10.1037/h0061626>
- Tong, T., Chignell, M., Lam, P., Tierney, M. C., & Lee, J. (2014). Designing serious games for cognitive assessment of the elderly. *Proceedings of the International Symposium on Human Factors*

- and *Ergonomics in Health Care*, 3(1), 28–35. <https://doi.org/10.1177/2327857914031004>
- Tong, T., Chignell, M., Tierney, M. C., Lee, J., & Tong, T. (2016). A serious game for clinical assessment of cognitive status: Validation study. *JMIR Serious Games*, 4(1), 1–14. <https://doi.org/10.2196/games.5006>
- Tuena, C., Mancuso, V., Stramba-Badiale, C., Pedroli, E., Stramba-Badiale, M., Riva, G., & Repetto, C. (2021). Egocentric and allocentric spatial memory in mild cognitive impairment with real-world and virtual navigation tasks: A systematic review. *Journal of Alzheimer's Disease*, 79(1), 95–116. <https://doi.org/10.3233/JAD-201017>
- Vasilyeva, M. (2005). Spatial cognition and perception. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 3, pp. 591–597). Elsevier Academic Press.
- Vasser, M., Kängsepp, M., Magomedkerimov, M., Kilvits, K., Stafinjak, V., Kivisik, T., Vicente, R., & Aru, J. (2017). VREX: An open-source toolbox for creating 3D virtual reality experiments. *BMC Psychology*, 5(1), 4. <https://doi.org/10.1186/s40359-017-0173-4>
- Ventura, M., Shute, V., Wright, T., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2013.00852>
- Vingerhoets, G., Lannoo, E., & Bauwens, S. (1996). Analysis of the money road-map test performance in normal and brain-damaged subjects. *Archives of Clinical Neuropsychology*, 11(1), 9. [https://doi.org/10.1016/0887-6177\(95\)00055-0](https://doi.org/10.1016/0887-6177(95)00055-0)
- Vlcek, K., & Laczó, J. (2014). Neural correlates of spatial navigation changes in mild cognitive impairment and Alzheimer's disease. *Frontiers in Behavioral Neuroscience*, 8, 1–6. <https://doi.org/10.3389/fnbeh.2014.00089>
- Wiener, J. M., Carroll, D., Moeller, S., Bibi, I., Ivanova, D., Allen, P., & Wolbers, T. (2020). A novel virtual-reality-based route-learning test suite: Assessing the effects of cognitive aging on navigation. *Behavior Research Methods*, 52(2), 630–640. <https://doi.org/10.3758/s13428-019-01264-8>
- Winters, M., & Weir, A. (2017). Grey matters; on the importance of publication bias in systematic reviews. *British Journal of Sports Medicine*, 51(6), 488–489. <https://doi.org/10.1136/bjsports-2016-096679>
- Zen, D., Byagowi, A., Garcia, M., Kelly, D., Lithgow, B., & Moussavi, Z. (2013, November). The perceived orientation in people with and without Alzheimer's. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 460–463). IEEE. <https://doi.org/10.1109/ner.2013.6695971>
- Zheng, J. M., Chan, K. W., & Gibson, I. (1998). Virtual reality. *IEEE Potentials*, 17(2), 20–23. <https://doi.org/10.1109/45.666641>
- Zucchella, C., Sinforiani, E., Tassorelli, C., Cavallini, E., Tost-Pardell, D., Grau, S., Pazzi, S., Puricelli, S., Bernini, S., Bottiroli, S., Vecchi, T., Sandrini, G., & Nappi, G. (2014). Serious games for screening pre-dementia conditions: From virtuality to reality? A pilot project. *Functional Neurology*, 29(3), 153–158. <https://doi.org/10.11138/FNeur/2014.29.3.153>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.