

A systematic review of using immersive virtual reality to assess executive function

Rebecca Kirkham, Lars Kooijman, Lucy Albertella, Dan Myles, Murat Yücel,
Kristian Rotaru

Submitted to: JMIR Serious Games
on: July 03, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	42
Figures	43
Figure 1.....	44
Multimedia Appendixes	45
Multimedia Appendix 1.....	46
License to publish form (upload after acceptance)s	47
License to publish form (upload after acceptance) 0.....	47
License to publish form (upload after acceptance) 0.....	47
License to publish form (upload after acceptance) 0.....	47

A systematic review of using immersive virtual reality to assess executive function

Rebecca Kirkham¹ BPsy(Hons); Lars Kooijman²; Lucy Albertella¹; Dan Myles¹; Murat Yücel^{1*}; Kristian Rotaru^{3*}

¹BrainPark Turner Institute for Brain and Mental Health, School of Psychological Sciences and Monash Biomedical Imaging Facility Monash University Clayton AU

²Institute for Intelligent Systems Research and Innovation Deakin University Geelong AU

³Monash Business School Monash University Caulfield AU

*these authors contributed equally

Corresponding Author:

Rebecca Kirkham BPsy(Hons)

BrainPark

Turner Institute for Brain and Mental Health, School of Psychological Sciences and Monash Biomedical Imaging Facility

Monash University

770 Blackburn road

Clayton

AU

Abstract

Background: Neuropsychological assessments have traditionally focused on executive functioning (EF) due to the significant role it plays in everyday life and its association with a variety of mental disorders. Although traditional assessment methods of EF are well established, they lack ecological validity and are limited to evaluating a single cognitive process. While these traditional EF assessment tools may be appropriate for clinical populations, they provide less information about EF in healthy individuals. Recognizing these limitations, there has been a growing interest in Virtual Reality (VR)-based assessments of EF due to their potential to increase test sensitivity, ecological validity, and accessibility to neuropsychological assessment.

Objective: This systematic review aims to explore the literature on VR assessments of EF. Specifically, we intended to identify: (1) the EF components being assessed, (2) how these assessments are validated, and (3) the monitoring strategies for potential adverse (cybersickness) and beneficial (immersion) effects.

Methods: EBSCOHost, Scopus, and Web of Science (WoS) were searched using keywords that reflected the main themes of Virtual Reality, Neuropsychological Tests and Executive Function. The search retrieved 555 unique articles, of which 19 met the inclusion criteria. Articles had to be an English written peer-reviewed manuscript published after 2013 that detailed an empirical, clinical or proof of concept study where a virtual environment using a head-mounted display was employed to assess EF in an adult population.

Results: The reviewed articles comprised a range of EF constructs, including inhibitory control, cognitive flexibility, working memory, planning, and attention. Although VR paradigms were typically validated against gold-standard traditional tasks, some studies did not report a-priori planned correlations, others did not detail the target EF constructs to be assessed by the VR paradigm, and there was often incomplete reporting of results. Further, cybersickness was assessed by only four of the 19 studies.

Conclusions: The authors propose several recommendations to improve the theory and practice of VR assessments of EF. A key conclusion is the critical need to address the methodological and psychometric properties during the development of these assessments to ensure their validity and reliability. Furthermore, we suggest an increased focus on potential adverse effects such as cybersickness. Future research should consider the integration of biosensors into VR systems, and the potential of VR in assessing spatial navigation. While VR assessments show promising potential, their adoption must be systematic and validated to ensure their utility in real-world applications.

(JMIR Preprints 03/07/2023:50282)

DOI: <https://doi.org/10.2196/preprints.50282>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>



Original Manuscript

A systematic review of using immersive virtual reality to assess executive function

Abstract

Background: Neuropsychological assessments traditionally include tests on executive functioning (EF) due to its critical role in daily activities and link to mental disorders. Established traditional EF assessments, although robust, lack ecological validity and are limited to single cognitive processes. These methods, suitable for clinical populations, are less informative for EF in healthy individuals. With these limitations in mind, immersive Virtual Reality (VR)-based assessments of EF have garnered interest for their potential to heighten test sensitivity, ecological validity, and neuropsychological assessment accessibility.

Objective: This systematic review aims to explore the literature on immersive VR assessments of EF, focusing on (1) EF components being assessed, (2) how these assessments are validated, and (3) strategies for monitoring potential adverse (cybersickness) and beneficial (immersion) effects.

Methods: EBSCOHost, Scopus, and Web of Science were searched using keywords that reflected the main themes of Virtual Reality, Neuropsychological Tests and Executive Function. The search retrieved 555 unique articles, of which 19 met the inclusion criteria. Articles had to be an English written peer-reviewed manuscript published after 2013 that detailed an empirical, clinical, or proof of concept study where a virtual environment using a head-mounted display was employed to assess EF in an adult population.

Results: Reviewed studies encompassed EF and associated higher-order cognitive functions such as inhibitory control, cognitive flexibility, working memory, planning, and attention. VR assessments commonly underwent validation against gold-standard traditional tasks. However, discrepancies were observed with some lacking reported a-priori planned correlations, omitting detailed descriptions of the EF constructs evaluated by the VR paradigms, and frequent incomplete result reporting. Notably, only four out of the 19 studies evaluated cybersickness.

Conclusions: Several recommendations are proposed to improve the theory and practice of immersive VR assessments of EF. Addressing the methodological and psychometric rigour is paramount for ensuring assessment validity and reliability. An augmented focus on monitoring potential adverse effects, particularly cybersickness, is warranted. Additionally, future studies should explore the integration of biosensors with VR systems and the capabilities of VR in the context of spatial navigation assessments. VR assessments hold considerable promise, but their systematic and validated implementation is essential to ensure their utility in real-world applications.

Keywords: Virtual Reality; Executive Functioning; Neuropsychological assessment; Systematic review

Introduction

Executive functioning (EF) has long been a focus of neuropsychological assessment due to the significant role it plays in everyday functioning. EF is an umbrella term for higher-order cognitive skills used to control and coordinate a wide range of mental processes and everyday behaviours [1–5], including “... mentally playing with ideas; taking the time to think before acting; meeting novel, unanticipated challenges; resisting temptations; and staying focused” ([6] p. 135). While a universally accepted definition of EF does not exist [5], there is agreement on the attributes of three core executive functions: inhibition, cognitive flexibility, and working memory [2,4,6]. These core executive functions support other higher-order executive functions such as reasoning, planning, and problem-solving [6–8]. Because EF impairment has been linked to a variety of mental disorders [9] it

is often considered a transdiagnostic risk factor [10].

Although traditional methods used to assess EF are popular [11,12] and well validated [13], they have been criticised for their lack of ecological validity [14,15]. Ecological validity, within the scope of this study, is defined as the “functional and predictive relationship between the person’s performance on a set of neuropsychological tests and the person’s behavior in a variety of real world settings” ([16] p. 15). Specifically, we interpret ecological validity as comprising two principal components: representativeness – the degree to which a neuropsychological test mirrors the demands of a person’s daily living activities it aims to evaluate [17], sometimes referred to as verisimilitude [18]; and generalizability – the extent to which test performance predicts an individual’s functioning in their daily living activities [17], also known as veridicality [18].

Traditional assessments tend to take a ‘construct-led’ approach, with each test intended to isolate a single cognitive process in an abstract measure. This process of abstraction may limit the ecological validity of the measure by resulting in poor alignment between the test outcomes and real-world functioning. In turn, this produces a large amount of variance in EF that is unaccounted for by traditional tasks. For example, Chaytor and colleagues [19] noted that traditional executive functioning tests accounted for only 18% to 20% of the variance in the everyday executive ability of participants. This lack of explained variance may be contributed to by the nature of the testing environment, the constructs assessed in isolation, the participant’s affective state and the compensatory strategies available to the participant [19]. A related methodological issue, known as the ‘task impurity problem’ [4,20], indicates that the score on an EF task usually reflects not only the systematic variance attributable to the specific aspect of EF targeted by that task, but also the 1) systematic variance across multiple types of EF tasks, 2) systematic variance attributable to non-EF aspects of the task, and 3) non-systematic (error) variance (see [10] for a detailed review). Outside the testing environment, the process of making a decision or planning and eliciting goal-directed behaviour in everyday life is often highly dynamic and influenced by numerous internal and external factors (e.g., [13,14]). Therefore, an ecologically valid assessment tool will need to include relevant contextual, dynamic, and multidimensional features, such as affect and physiological state, that traditional assessments cannot.

Further, while traditional EF assessment tools may be appropriate for clinical populations, they generate less information about functioning in relatively healthy individuals. For example, the trail-making task has routinely been administered as a neuropsychological assessment of driving performance. While some studies demonstrate a relationship between the two [21,22], others show no relationship [23], particularly in non-clinical populations [24,25]. Thus, while traditional tools are adequate for detecting more severe EF impairment, they are less effective in detecting subtle changes in EF functioning and early decline. Increased test sensitivity to detect subtle intraindividual changes may enable better detection of the prodromal stage of cognitive decline. Early detection is important because it enables early intervention, which may in turn improve prognosis. For example, sensitive detection can identify the prodromal stages of Alzheimer’s Disease in seemingly healthy individuals [26] and mild cognitive decline up to 12 years prior to clinical diagnosis [27]. Similarly, in a situation where an individual requires a capacity assessment for an activity, traditional assessments may have limited utility for non-clinical populations. The triangulation of multiple data sources, such as biosensors, may increase sensitivity to better identify subtle changes in capacity.

To address the shortcomings of poor ecological validity and test sensitivity, research into psychological assessment has begun to investigate Virtual Reality (VR) technology as a means of providing a more naturalistic environment for assessing EF in clinical neuropsychological assessments. VR enables the development of custom-designed, simulated environments that can

replicate real-life environments, potentially increasing its ecological validity though representativeness. In addition, VR could increase engagement [28,29], reduce test time, and better integrate data from biosensors with in-task events that facilitate assessment. The following sections will expand on these points and consider the importance of validating and assessing the reliability of VR for EF assessment.

Ecological validity and representative tests

There is an increasing emphasis on conducting EF assessments using tasks that resemble situations experienced in everyday life [30]. For example, the Multiple Errands Test [31] requires individuals to run errands in a real environment (e.g., a shopping centre). Empirical assessment of the MET has demonstrated its generalizability to daily function [32] and carer reports of daily functioning [33]. However, given that the MET is designed to be performed in real-life locations, it is impractical for routine administration by clinicians [34,35] and is susceptible to the variable features of real-world environments which are outside experimental control. VR can mitigate these difficulties by maintaining the real-world environment without requiring travel, whilst enabling fine-tuned control and uniform presentation of environmental characteristics [36]. Several studies [37–39] have investigated and developed platforms for this purpose, commonly known as the Virtual Multiple Errands Test.

Engagement

VR holds the potential to enhance individual engagement more effectively than traditional pencil-and-paper or computerized tasks, by offering a fully immersive experience [40]. Recognized as a crucial aspect of cognitive assessment, engagement can be improved through gamification, thereby improving task performance [41]. ‘Serious games’, defined as games intended for a variety of serious purposes, such as training, learning, stimulation, or cognitive assessment [42], have been shown to be more engaging than non-gamified tasks [43–45]. The unique immersive environment of VR captures increased attention, leading to reduced average response times and response time variability [46]. Notably, recent studies using EEG-based metrics show greater attention elicited in immersive VR paradigms compared to 2D computerized assessment. [46]. This heightened immersion and engagement in VR may enhance the reliability of measures by capturing a more accurate representation of an individual’s best effort.

Cybersickness

Despite their increased engagement, VR paradigms have the potential to induce cybersickness which can threaten the validity of the paradigm. Cybersickness (i.e., dizziness, vertigo) is akin to motion sickness but occurs in response to exposure to VR [47]. Previous research suggests that there is a negative relationship between cybersickness and cognitive abilities. For example, Nalivaiko and colleagues [47] found that reaction times were moderately correlated ($r = 0.5$, $P = 0.006$) with subjective ratings of nausea. Similarly, Sepich and colleagues [48] found that participants’ accuracy on n-back task performance was weakly to moderately negatively correlated ($r = -0.32$, $P = 0.002$) with subjective cybersickness ratings. Therefore, there is reasonable concern that the potential benefits of engagement and ecological validity may be compromised if participants experience cybersickness.

Validity, reliability, and sensitivity

Arguably, the biggest threat to the utility of VR platforms is that many studies do not document their validity and reliability. A meta-analysis has shown that VR assessment tools are moderately sensitive to cognitive impairment across neurodevelopmental, mental health, and neurological disorders [49], demonstrating their promising application in clinical settings. Borgnis and colleagues [50] reviewed the VR-based tools for EF assessment that are currently available, illustrating the plethora of platforms developing in this field. The works by Negut and colleagues [49] and Borgnis and

colleagues [50] highlight the utility of VR assessment tools to detect dysfunction and present the various tools in literature created to investigate EF. Kim and colleagues [51] provided an overview of the research trends using VR for neuropsychological tests, and documented the cognitive functions assessed by each study. However, to the best of our knowledge, there is no overview or examination of the of the psychometric properties of these VR tools or how their psychometric properties are being evaluated.

Typically, novel measures and assessments are validated against current gold-standard tasks for concurrent validity [52]. Concurrent validity can be a reliable means of determining whether two assessments measure the same construct. However, concurrent validity can also occur when two tests contain the same problems, such as inaccurately measuring a particular construct in the same way. Sequentially, many VR tasks are being created from a 'function-led' perspective, but are validated against 'construct-led' tasks [53,54]. Given their different approaches, function-led and construct-led assessments should be validated in different ways, or at least, with several validation approaches. If function-led VR assessments improve upon the validity of current assessment methods, validation techniques may also need to go beyond comparisons with traditional assessments. For example, function-led VR assessments may be better validated against additional alternate methods, such as carer reports, real-life performance (e.g., self-care, residence, transportation, employment), and diagnostic trajectory [49], as opposed to validating through traditional (construct-led) assessment. Without incorporating tests of ecological validity, the potential advantages of VR may go unrecognised. Given the increasingly rapid development of VR neuropsychological assessments, it will be imperative to maintain high validation standards for these tools (e.g., [55]).

Establishing the reliability of novel VR EF assessments is also critical to the integrity of the outcomes. Reliability ensures that the measure yields consistent and repeatable results, a foundational element for test validity. Consequently, both reliability and validity ought to be evaluated for each measurement tool. Test-retest reliability, confirming consistency over time, should be accompanied by the interval between assessments and the correlation of results. Internal consistency, typically measured by Cronbach's alpha, should also be reported for each target construct or domain of assessment. Importantly, for immersive VR EF assessments that evaluate multiple EF constructs, it is essential to report the alpha for each distinct construct, rather than a collective coefficient. This is because the coefficient is intended to evaluate item consistency within a scale measuring a single construct, applying it across disparate constructs could be confusing and potentially misleading.

Consistency of terminology

Finally, to ensure psychometric precision and build on previous research, EF assessment paradigms must adopt consistent terminology for their target assessment constructs. The field of EF, although of significant interest to both researchers and clinicians, is marked by varied terminology for identical constructs. This issue, longstanding in EF research (see Suchy [5], for a review see Baggetta & Alexander [56]) presents challenges to VR in the EF assessment field. For instance, inconsistent terminology hinders the synthesis of research findings. Diverse labels such as 'impulsivity' and 'impulse control' might, upon examination, refer to the same underlying construct. Consequently, researchers aiming to extend the literature on 'impulsivity' might overlook pertinent studies or exclude valuable references due to terminological discrepancies.

This literature review seeks to examine and discuss the development of the VR tools used to assess EF with a specific focus on evaluating their psychometric properties. The studies selected for inclusion in this review are those that have developed assessment tools for EF either holistically, or

in part. The aims of this review are to 1) determine the components of EF assessed by VR paradigms, 2) investigate the methods used to validate VR assessments, and 3) explore the frequency and efficacy of reporting participants' immersion and engagement in VR for EF assessment.

Methods

Our review methodology followed the PRISMA statement [57]. In line with literature, EF was defined as a set of executive functions, including inhibition, cognitive flexibility, and working memory [2,4,6], which support other higher-order executive functions, such as reasoning, planning, and problem solving [6,8].

Inclusion Criteria

Prior to conducting the literature search, the inclusion criteria were established. First, only those peer-reviewed articles and conference proceedings (complete manuscripts) written in English were included. Second, articles were included which detailed either an empirical, clinical, or proof of concept study where an immersive virtual environment (i.e., using a head-mounted display, not a 2D computer screen) was reported to investigate EF or higher-order cognition broadly, or examined EF via a selection of one or more sub-constructs (e.g., inhibitory control, working memory). Finally, only those articles with an adult participant population published after 2013 were included. This temporal limit was based on the release date of the Oculus Rift Developer Kit 1, since it was one of the first accessible products for public use of VR. Articles were identified through EBSCOHost, Scopus, and Web of Science (WoS) citation databases. Scopus and WoS were chosen due to their prominence as citation databases [58]. To compensate for the bias towards Engineering and Natural Science articles found through Scopus and WoS [59], EBSCOHost was searched for articles published in fields such as (clinical) psychology and medicine.

Search Strategy

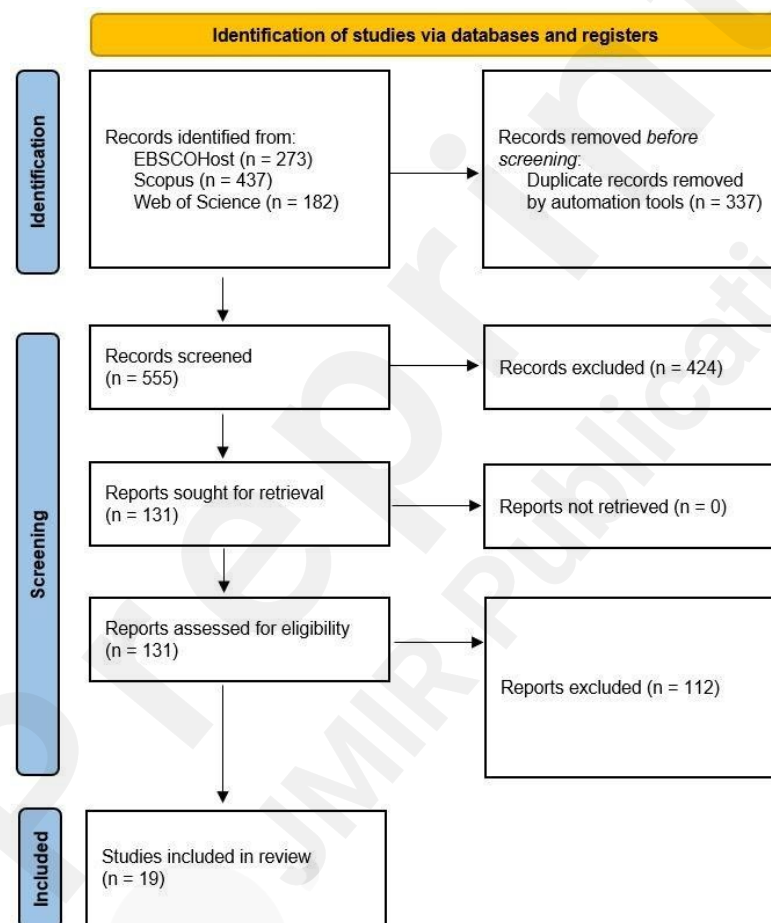
Keywords were developed by identifying three main components which the relevant literature should include. The three components were based on "Virtual Reality" and "Neuropsychological Tests" and "Executive Function". It was decided not to search specific components of executive functioning due to the lack of consensus in the field regarding its components. Rather, it was assumed that if a study was addressing EF or a component of EF, it would include "executive functioning" as a keyword in either the title, abstract, or keywords. Other reviews looking broadly at VR paradigms have used similar search strategies [49].

In the current study, key terms were developed by identifying synonyms for key components and concatenating them with 'AND' Boolean operators. The final keywords used for the search were: ((("virtual" OR "artificial" OR "simulated") AND ("realit*" OR "world" OR "environment"))) AND ((neuropsych* OR function* OR cognit*) AND ((executive AND function*) OR (high* AND order AND cognit*))) AND (assessment)).

The systematic literature-search process is shown in Figure 1. Literature queries made through EBSCOHost were limited to the following databases: Academic Search Complete, AgeLine, AMED, Applied Science and Technology Source, CINAHL, E-Journals, Health Source Consumer and Nursing/Academic Edition, MEDLINE, Mental Measurements Yearbook, Psychology and Behavioral Sciences Collection, and all variations of APA databases. Furthermore, for the search, three topic fields (i.e., Title, Abstract, and Subject Terms) were used to paste the keywords. The three topic fields were concatenated with 'OR' Boolean operators. As a result, 273 items were identified through our query using EBSCOHost. Using the Scopus database, we implemented a basic search in the article title, abstract or keywords using the keywords. No additional limiters were applied, and 437 items were identified. Our search in Web of Science (WoS) included all databases, and the advanced search method was used wherein keyword searches in the article title, abstract and

keyword topic fields were concatenated by 'OR' Boolean operators (i.e., Title = (keywords) OR Abstract = (keywords) OR Keywords = (keywords)). Through WoS, 182 items were identified. Results for each database were exported into the Covidence systematic review management platform [60], which filtered out 337 duplicates. A total of 555 unique articles remained. All abstracts were screened independently by the first author and the senior author to determine whether the contents met the inclusion criteria. Subsequently, 424 were deemed irrelevant. The final 131 articles had their full text screened by the same authors and 19 articles met the inclusion criteria. Any disagreement was discussed by the first (RK), second (LK), and the senior author (KR).

Figure 1. Systematic review process and results from literature searches in EBSCOHost, Scopus and Web of Science databases



Data extraction

The first and second author completed the data extraction process by manually reviewing each manuscript; data items (see below) were recorded using Microsoft Excel.

Data items

A qualitative evaluation of the studies included in the review was performed, meaning that the content of each manuscript was assessed based on the reported target construct(s) relevant to EF, and the extent to which the reported virtual task related to the assessment of the target construct(s). Studies were categorized based on the construct they targeted through their VR paradigm, as reported by the authors of the respective articles. If multiple constructs were assessed in a single study, the study was included for each construct. No inferences were made about which cognitive construct(s) was assessed based on the tasks that were reported in the manuscript. For example, if a

study indicated only that they used a VR version of the Stroop test but did not disclose which construct it was assessing with this test, the study was not categorized under inhibitory control or cognitive flexibility but under the general factor 'executive functioning'. Furthermore, it was denoted whether studies clearly identified which validation tasks were correlated to VR paradigm measures or if the correlations were implied. For instance, if the study directly stated a-priori that they hypothesised a correlation between a VR task measuring inhibition and a validation task, like the Stroop Task, this was recognized as providing explicit validation for inhibition. Conversely, if a study indicated that participants completed the Stroop task, which assessed inhibition and processing speed, and mentioned that the VR paradigm evaluated inhibition, it was considered to provide implicit validation for inhibition. Studies incorporating measures for immersion, engagement, or cybersickness were also documented.

Results

General executive functioning

Seven of the reviewed studies assessed EF in general, meaning the respective authors of these studies did not explicitly state which subconstruct of EF was targeted with the VR task. Table 1 shows which validation tasks were used in each study to measure EF.

Banville and colleagues [61] immersed participants in a Virtual Multitasking Test (VMT), which was in principle designed to measure prospective memory and executive functions by having participants perform multiple tasks in a virtual apartment. However, this study reported specifically on the task where participants had to store groceries as fast as possible, while also being attentive to other tasks, such as answering the phone or closing a window. Although the authors hypothesized that VMT scores would be correlated to neuropsychological assessment, such as mental flexibility, planning, and inhibition, it was not explicitly stated which metric of the VMT would be correlated to which neuropsychological assessment. Nonetheless, the authors identified that grocery storing time was correlated to the break of rules score from Six Element Test (SET; $r(19) = -0.49$, $P = 0.04$; P -value as reported in manuscript). Furthermore, the number of errors in storing fruit and vegetable was found to correlate with the perseveration score of the Zoo Map Test (ZMT; $r(20) = 0.53$, $P = 0.02$; P -value as reported in manuscript) and reading speed during the second condition of the Stroop Test (ST; $r(20) = 0.44$, $P = 0.05$; P -value as reported in manuscript).

Davidson and colleagues [62] immersed participants in a parking simulator and a chemistry lab where participants had to either park a vehicle, sort chairs, or locate items. Prior to immersion, participants completed the Stroop Test (ST) and the trail making test A and B (TMT-A/B). The authors identified that completion time of the second level (Kendall's tau, $\tau = -0.32$, $P = 0.01$; P -value as reported in manuscript) and the number of levels completed of the parking simulator ($\tau = 0.43$, $P < 0.01$; P -value as reported in manuscript) were correlated to participants' performance on the ST. Additionally, ST was correlated to seating arrangement metrics, such as time to place first stool ($\tau = -0.33$, $P = 0.01$; P -value as reported in manuscript) and number of stools placed ($\tau = 0.33$, $P = 0.02$; P -value as reported in manuscript), as well as correlated to time to locate the first item in the chemistry lab ($\tau = -0.37$, $P = 0.01$; P -value as reported in manuscript). Correlations between TMT-A/B and, for example, number of completed parking levels ($\tau = -0.49$, $P < 0.01$; P -value as reported in manuscript) or the number of items placed in the seating arrangement task in the chemistry lab ($\tau = -0.35$, $P = 0.01$; P -value as reported in manuscript) were reported. However, reporting was limited to significant correlations only and no a-priori expectation how performance on VR tasks and validation tasks were correlated was denoted in the study.

Miskowiak and colleagues [63] assessed executive functions by letting participants complete TMT-B, One-touch stockings of Cambridge mean choices to correct (CANTAB) and Verbal Fluency Test (VFT) version S and D. Performance of these tests were compared to participants' performance of a cooking task in VR. The authors hypothesized that the number of cooking tasks that were correctly

placed on a to-do list and the latency to solving the task would be VR-equivalent measures of executive functioning. The authors found that VR performance was correlated ($r(121) = 0.26$, $P = .004$) to executive functioning, which consisted of a correlation between the average performance on the VR subtasks and the average performance on the validation tasks. Correlations between the individual performance of VR and validation tasks were not reported in the manuscript.

Pallavicini and colleagues [64] had participants play the Audioshield dance game, which the authors hypothesised could be closely related to EF functions, such as inhibition and working memory. However, the authors correlated participants' performance at the Audioshield game to their performance on Trail-Making Task (TMT) version A and B which measures psychomotor speed (TMT-A) and mental flexibility (TMT-B). Nonetheless, the results showed that TMT performance was negatively correlated to Audioshield performance metrics.

Porffy and colleagues [65] had participants complete VStore, where the two tasks measured EF, namely the "Find" task and the "Coffee" task. Specifically, participants had to find 12 items from a list they previously memorised. Additionally, participants had to order a hot drink from the coffee shop after finding, bagging, and paying for the 12 remembered items they found in the store. Notably, the authors indicated that the two VR tasks also tapped into navigation (i.e., Find) and processing speed (i.e., Coffee). Furthermore, the Groton Maze Learning (GTML) from Cogstate was used to evaluate general EF which the participants completed prior to the VR task. Nonetheless, through their regression analysis, the authors identified that GTLM was not a predictor for the Find task ($B(SE) = 0.024 (0.029)$, $P = .11$; P -value as reported in manuscript) or Coffee task ($B(SE) = -0.003 (0.051)$, $P = .96$; P -value as reported in manuscript).

Tan and colleagues [66] had 100 participants complete 13 tasks in a virtual environment, which were designed to measure six cognitive domains, such as executive functioning and complex attention. Although differences in performance on VR tasks related to EF between age groups were found, no comparison was reported or made to a traditional neuropsychological assessment of EF or any subconstructs of EF.

Tsai and colleagues [67] immersed two participant groups into a virtual shopping environment, one group with mild cognitive impairments (MCI) and one control group. The VR tasks assessed participants' memory, EF, and calculation, by having participants memorise a shopping list, search for the listed items in the shop, and subsequently pay for them. The authors trained machine learning models on features extracted from the VR tasks to predict whether participants had MCI or were healthy controls, which was achieved with high accuracy. Nonetheless, no neuropsychological assessment of EF was reported as a validation for the VR tasks.

Table 1. Reviewed studies targeting general executive functioning, their respective validation task, author and year, and total number of studies examining general executive functioning

VR target construct	Validation Task	Validation	Authors	Number of studies examining construct
Executive functioning - general	Delis-Kaplan Executive Function System (D-KEFS) Trail Making Test (TMT AB) Stroop Task (ST) Modified version of the Six Element Test	Implicit	Banville and colleagues [61] ^a	7

(SET) Hanoi Tower Test (HTT) Zoo Map Test (ZMT)			
ST TMT AB	Implicit	Davison and colleagues [62] ^b	
TMT B One-touch stockings of Cambridge mean choices to correct (OTS, CANTAB) Verbal Fluency Test (S and D) (VFT)	Explicit	Miskowia k and colleagues [63]	
TMT AB	Explicit	Pallavicini and colleagues [64]	
Groton Maze Learning - Cogstate	Implicit	Porffy and colleagues [65]	
None specifically reported	N/A	Tan and colleagues [66]	
None specifically reported	N/A	Tsai and colleagues [67]	

^aBanville and colleagues [61]: The VR task was predominantly a sorting task for executive function assessment. The comparative assessments that validated this assessment have been detailed under 'executive function' broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.

^bDavison and colleagues [62]: The VR task was reported to assess executive function assessment. The comparative assessments that validated this assessment have been detailed under 'executive function' broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.

Targeted Constructs

The subsections below elaborate on the EF (sub-)constructs addressed by the studies under review. A range of correlation coefficients are reported in these studies; however, due to the lack of uniformity in result reporting, these coefficients are omitted from the current synthesis. Typically, the studies reported only significant correlations between metrics without presenting all potential correlations. Furthermore, only three studies specified an alpha level (i.e., 0.05), with other three studies indicating statistical significance at a *P*-value equal or smaller than 0.05. Four studies did not denote an alpha level but mentioned applying corrections for multiple comparisons, yet did not detail the adjusted alpha level. One study adopted Bayesian statistics, using a Bayesian factor over 10 for statistical inference. Nonetheless, the reviewed studies did not consistently clarify which VR tasks were validated against traditional tasks, hindering the construct validity for the various EF

components. Consequently, drawing consistent conclusions on how EF (sub-) constructs were evaluated was not feasible without inferring the nature of the tests and assessment paradigms.

Core Executive Functions

Inhibition

Of the three “core” executive functions, seven unique studies of the 19 included in our review investigated inhibitory control, interference control or impulsivity, either singly or combined. Table 2 details the respective validation task and target construct of each of these studies. For example, Chicchi Giglioli and colleagues [68] presented participants with six standardized tasks, three of which assessed inhibition (see Table 2), prior to administering a serious game wherein participants were required to perform tasks in outer space. Ten of the 36 possible correlations between measures for the standardized tasks and the serious game tasks were reported as statistically significant and ranged from weak ($.20 < r < .39$; relative P -values indicated in manuscript, e.g., $P < .05$) to strong ($.60 < r < .79$; relative p -values indicated in manuscript) positive. For example, the latency metric of the Dot Probe task correlated positively ($.35 < r < .54$; relative P -values indicated) with the latency metric of the three VR tasks aimed to measure inhibition, whereas no correlations were reported between the correct answer metric of the Dot Probe task and the correct answer metric of the three VR tasks aimed to measure inhibition. None of the metrics from the Stroop task correlated with the virtual task (requiring participants to fight aliens); however, the correct answer and latency metric of the Stroop task correlated with the virtual task (requiring participants to repair a valve).

Similarly, Chicchi Giglioli and colleagues [69] immersed participants in a virtual kitchen where participants had to cook different types of food. The activities were grouped in four subtasks of incremental difficulty, where in the third level inhibition was assessed by determining whether the right dressing was added using a go/no-go type paradigm. The authors detailed that the dot-probe, Go/No-Go, and Stroop tasks were utilized as standard tasks to assess inhibition. The unspecified metric of ‘correct dressing’ was shown to correlate well ($r = .527$, $P = .01$; relative P -value indicated in manuscript) with the correct answer metric of the Stroop task in one group, whereas in the second group, a moderate negative correlation ($r = -.486$, $P = .05$; relative P -value indicated in manuscript) was found between the execution time of the Tower of London task and the correct dressing metric. However, no other correlations between the virtual task metric and the traditional assessments that assessed inhibition were reported.

Table 2. Reviewed studies targeting inhibition, their respective validation task, author and year, and total number of studies examining the construct

VR target construct	Validation task	Validation	Authors	No. of studies
Inhibition / Inhibitory Control	Dot Probe Task (DPT) Go/No-Go Task (GNG) Stroop Task (ST)	Implicit	Chicchi Giglioli and colleagues [69]	6
	DPT GNG ST	Explicit	Chicchi Giglioli and colleagues [68]	

	GNG	Implicit	Marín-Morales and colleagues [70]	
	Continuous Performance Test (CPT)	Implicit	Voinescu and colleagues [71] ^b	
	None specifically reported	N/A	Parsons & Carlew [72]	
	ST	Implicit	Parsons & Barnett [73]	
Interference control	ST	Implicit	Marín-Morales and colleagues [70] ^a	3
	The Color Word Interference Test from the D-KEFS (CW-IT, D-KEFS) Automated Neuropsychological Assessment Metrics (ANAM) Stroop task	Implicit	Parsons & Carlew [72]	
	CW-IT, D-KEFS	Implicit	Parsons & Barnett [73]	
Impulsivity	None specifically reported	N/A	Chicchi Giglioli and colleagues [68]	1

^aMarín-Morales and colleagues [70]: The VR task involves 42 VR mini-games which assess various cognitive constructs. Four mini-games and their target constructs are documented and included in this table, however the comparative assessments were not provided, and an extensive list of all 42 mini-games was not provided.

^bVoinescu and colleagues [71]: Some traditional tasks listed were included for divergent validity and therefore have been omitted from the table.

Task index: ST = Stroop Task; DOT = Dot Probe Task; GNG = Go No-Go Task; CPT = Continuous performance Test [64]; CW-IT, D-KEFS = The Color Word Interference Test from the D-KEFS.

Marín-Morales and colleagues [70] had participants complete neuropsychological assessments, among which the Go/No-Go task, as well as 42 mini games in VR. An undisclosed set of variables from the minigames were used as predictors for measures of neuropsychological batteries. The mini-

game predictor variables were fed into different machine learning algorithms. The authors highlighted that games related to inhibition produced worse results compared to other games but did not report any results on inhibition. The authors did find that mini-game features of planning and attention could predict Go/No-Go hits proportions and mean time with 80 and 94% accuracy, respectively.

Parsons & Carlew [72] had participants perform the Stroop Task in a virtual classroom as well as complete a computerized and paper-and-pencil version of the task. The authors found that participants' performance was lower for colour naming and word reading in the VR paradigm compared to paper-and-pencil, but interference performance was better in the VR paradigm compared to paper-and-pencil. Similarly, Parsons & Barnett [73] had participants perform the Stroop Task in a virtual apartment as well as complete a computerized and paper-and-pencil version of the task. Here, the authors found that participants were more accurate in the Stroop task in the paper-and-pencil version compared to the VR paradigm.

Voinescu and colleagues [71] immersed participants in a virtual aquarium wherein they had to perform a variety of tasks. For example, participants had to respond when they saw a fish that was different from a clown fish or heard a different fish name than surgeonfish. After the VR aquarium, participants completed a variety of computerized tasks, amongst which the Continuous Performance Test (CPT) which was hypothesized to measure sustained attention and inhibition. The authors found weak to moderate ($.22 < r < .49$, relative P -values indicated, e.g., $P < .05$) correlations between CPT measures and VR measures.

Working memory

Working memory was investigated in four studies [63,65,70,74]. Table 3 details the respective validation task and target construct of each of these studies. The working memory component from the Marín-Morales and colleagues [70] study included a mini game wherein participants had to recall the ingredients on a recipe seen prior to the mini game and collect from a range of options only those ingredients found in a recipe. However, no correlates to neuropsychological tasks were presented. Miskowiak and colleagues [63] compared their VR paradigm with a traditional task that assessed working memory. In this study, participants were instructed to plan and cook a meal in a virtual kitchen. Performance metrics, such as the number of drawers opened and the latency until the task was completed, were used to assess working memory and were correlated to metrics from the traditional tasks such as WAIS Letter-number sequencing. The authors reported a significant positive correlation ($r(121) = .31$, $P = .001$) between the virtual task metrics and the traditional task metrics that evaluated working memory.

Porffy and colleagues [65] asked participants to operate a virtual store, whereby the working memory component was assessed at the 'Pay' step where participants had to select and pay for their items at a self-checkout machine, providing the exact amount. The authors specified that the reaction time on the one-back task and the accuracy of performance on the two-back task were metrics from traditional tasks used to assess working memory. By means of linear regression, the authors found that performance on the two-back task was negatively associated ($B(SE) = -0.085 (0.042)$, $P = .047$) to participants' performance on the pay step.

Robitaille and colleagues [74] assessed working memory during their simultaneous cognitive tasks where participants had to both recognise faces in windows that were previously declared as "hostile" or "non-hostile" and complete a navigation task. However, no correlations between traditional tasks and the virtual tasks were reported.

Table 3. Reviewed studies targeting working memory, their respective validation task, author and year, and total number of studies examining the construct

VR target construct	Validation task	Validation	Authors	No. of studies
Working memory	Wechsler Adult Intelligence Scale – WAIS-IV, The Working Memory Index (Digit Span and Arithmetic)	Implicit	Marín-Morales and colleagues [70] ^a	4
	WAIS-III letter-number sequencing (LNS, WAIS-III) Spatial working memory from CANTAB (SWM, CANTAB), error and strategy	Explicit	Miskowick and colleagues [63]	
	One-back and two-back – Cogstate	Implicit	Porffy and colleagues [65]	
	None specifically reported	N/A	Robitaille and colleagues [74] ^b	

^aMarín-Morales and colleagues [70]: The VR task involves 42 VR mini-games which assess various cognitive constructs. Four mini-games and their target constructs are documented and included in this table, however the comparative assessments were not provided, and an extensive list of all 42 mini-games not provided.

^bRobitaille and colleagues [74]: Used a VR paradigm with avatars to trial a dual task walking protocol

Cognitive flexibility

One study by Chicchi Giglioli and colleagues [68] investigated cognitive flexibility (termed “cognitive shifting” in the paper) through three virtual tasks. The authors specified that the trail-making task was used as a traditional task to assess cognitive flexibility as a comparator for the first virtual task (CF1, cultivating food), and the Wisconsin card sorting test was used as a traditional task to evaluate cognitive flexibility as a comparator for the other two virtual tasks (CF2, growing plants; CF3, fuelling a turbine). The total time metric of the first virtual task correlated positively with the total time of TMT-B ($r = .396$, $P < .01$; P -value as reported in the manuscript) and multiple metrics of virtual task 2 and 3 correlated with performance metrics of the Wisconsin card sorting task.

Higher-order Executive Functions

Planning

Five studies [62,68,69,75,76] identified planning as a target construct in their VR paradigms. Table 4 details the respective validation task and target construct of each of these studies. The VR environment created by Chicchi Giglioli and colleagues [69] used a cooking task with four levels of difficulty. In the three more difficult levels, planning was required to complete tasks as two burners were used. There was no clearly specified metric for the virtual task that was used to evaluate planning, but the authors specified that the Tower of London task was used as a traditional

assessment to evaluate planning. A variety of virtual task metrics, such as total time to complete a difficulty level, were shown to correlate with various Tower of London metrics.

In another study, Chicchi Giglioli and colleagues [68] used a VR paradigm based on an outer-space environment. The paradigm contained eight tasks, one of which assessed planning ability (Task 7). The authors stated that the Tower of London task was the traditional assessment tool used to evaluate planning, and explained that the total score, initial time, and execution time of the virtual task were the outcome metrics. Moderate positive correlations were found between the execution time of the virtual and the Tower of London task ($r = .463$, $P < .01$; P -value as reported in the manuscript), and the initial time of the virtual task and the total time of the Tower of London task ($r = .372$, $P < .05$). Furthermore, the virtual task correlated with some metrics of other traditional assessments used to assess planning ability, although these were not specified a-priori.

Both Kourtesis and colleagues [76] and Kourtesis and MacPherson [75] used the same VR environment based on a variety of everyday tasks. One task assessing planning ability required participants to draw their route around the city (e.g., visiting the bakery, supermarket, library, and returning home) on a 3D interactive board. Kourtesis and colleagues [76] explained that the Key Search task from the BADS was used as a traditional measure to assess planning and found a strong positive correlation between the traditional task and the virtual task ($r = .80$, Bayes Factor = $4.65e^8$). Furthermore, Kourtesis and MacPherson [75] noted in their results that planning explained a substantial 12% ($P = .030$) of variance in time-based prospective memory which was required in 10 of 17 tasks.

Davison and colleagues [62] assessed planning ability using a task involving the arrangement of a table and chair. However, they did not explicitly mention the traditional task that was used to evaluate planning. Various correlations between the performance metrics for the virtual task and the traditional tasks were reported. For example, the performance on the Stroop colour word task was negatively related to the time participants took to place a blue chair in the seating arrangement task (Kendall's tau, $\tau = -0.39$, $P = .01$; P -value as reported in the manuscript).

Table 4. Reviewed studies targeting planning, their respective validation task, author and year, and total number of studies examining each construct

VR target construct	Validation task	Validation	Authors	No. of studies
Planning	Tower of London-Drexler test (TOL-DX)	Implicit	Chicchi Giglioli and colleagues [69]	5
	Tower of London (TOL)	Explicit	Chicchi Giglioli and colleagues [68]	
	None specifically reported	N/A	Davison and colleagues	

			[62] ^a	
	The Key Search task from the BADS (Wilson and colleagues, 1996) [69]	Explicit	Kourtesis and colleagues [76]	
	None specifically reported	N/A	Kourtesis & MacPherson [75]	

^aDavison and colleagues [62]: The VR task was reported to assess executive function assessment. The comparative assessments that validated this assessment have been detailed under 'executive function' broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.

Other domains

Several studies examined domains of functioning that did not align with the EF definition used in this review. Broadly, these domains fell under categories of memory, attention, processing, task performance, and a variety of other uncategorized sub-constructs. Since the literature (e.g., [1] [2,4,6]) does not relate these broad domains to executive functioning, they are not discussed further but are denoted in Table 5.

Table 5. Reviewed studies targeting constructs unrelated to executive functioning, their respective validation task, author and year, and total number of studies examining each construct

VR target construct	Validation task	Validation	Authors	No. of studies
Memory				11
Memory (general)	None specifically reported	N/A	Tsai and colleagues [67]	1
Verbal Memory and Verbal Learning	Rey Auditory Verbal Learning Test (RAVLT) subtests: IV total, Immediate recall, delayed recall, recognition)	Explicit	Miskowick and colleagues [63]	2
	International Shopping List Task - Cogstate (Verbal learning)	Implicit	Porffy and colleagues [65]	
Prospective Memory	None specifically reported	N/A	Banville and colleagues [61] ^a	4

	Cambridge Prospective Memory Test (CAMPROMPT; Wilson, 2005) [71]	Explicit	Kourtesis and colleagues [76] ^b	
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
	California Verbal Learning Test - Second Edition (CVLT-II; Delis et al., [73])	Implicit	Parsons & McMahan [53]	
Episodic Memory	Rivermead Behavioral Memory Test-III (RBMT-III; Wilson et al., [74])	Explicit	Kourtesis and colleagues [76] ^b	3
	CVLT-II	Implicit	Parsons & McMahan [53]	
Immediate recognition	Rivermead Behavioral Memory Test-III (RBMT-III; Wilson et al., [74])	Explicit	Kourtesis and colleagues [76]	2
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Delayed Recognition	Rivermead Behavioral Memory Test-III (RBMT-III; Wilson et al., [74])	Explicit	Kourtesis and colleagues [76] ^b	2
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Attention				13
General attention	DPT GNG ST	Implicit	Chicchi Giglioli and colleagues . [69]	4
	DPT	Explicit	Chicchi	

	GNG ST TMT AB		Giglioli and colleagues [68]	
	Dot Probe Task - selective attention Go/NoGo Task - sustained attention Stroop Test - selective attention Trail Making Task - visual attention	Implicit	Marín-Morales and colleagues [70] ^c	
	Rapid Visual Processing from CANTAB (RVP, CANTAB) , accuracy and latency RBANS digit span (RBANS-DS)	Explicit	Miskowick and colleagues [63]	
Divided attention	None specifically reported	N/A	Robitaille and colleagues [74] ^d	2
	Color Trails Test -Trails B (CTT-B; D'Elia et al., 1996) [75]	Explicit	Wilf and colleagues [54]	
Complex attention	None specifically reported	N/A	Tan and colleagues [66]	1
Selective visual attention	The Map task from the TEA	Explicit	Kourtesis and colleagues [76] ^b	2
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Selective auditory attention	The Elevator Counting with Distraction task of the TEA	Explicit	Kourtesis and colleagues [76] ^b	2
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Sustained visual	CTT-A	Explicit	Wilf and colleagues	1

attention			[54]	
Visuospatial attention	The Ruff 2 and 7 Selective Attention Test.	Explicit	Kourtesis and colleagues [76] ^b	2
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Sustained attention	CPT	Implicit	Voinescu and colleagues [71]	1
Processing				3
Processing Speed	Wechsler Adult Intelligence Scale - WAIS-IV, The Processing Speed Index (Symbol Search and Coding)	Implicit	Marín-Morales and colleagues [70] ^c	3
	RBANS coding test (RBANS-CT) TMT-A	Explicit	Miskowick and colleagues [63]	
	Detection Task - Cogstate	Implicit	Porffy and colleagues [65]	
Task Performance				4
Dual Task	TMT-AB	Implicit	Chicchi Giglioli and colleagues [69]	1
Multitask	Modified version of the SET	Implicit	Banville and colleagues [61] ^a	3
	CTT	Explicit	Kourtesis and colleagues	

			[76] ^b	
	None specifically reported	N/A	Kourtesis & MacPherson [75]	
Uncategorised				12
Visual perception	None specifically reported	N/A	Marín-Morales and colleagues [70] ^c	1
Verbal learning	RAVLT subtests: IV total, Immediate recall, delayed recall, recognition	Explicit	Miskowick and colleagues [63]	2
	International Shopping List Task - Cogstate	Implicit	Porffy and colleagues [65]	
Navigation	None specifically reported	N/A	Porffy and colleagues [65]	2
	None specifically reported	N/A	Robitaille and colleagues [74] ^d	
Associate learning	Continuous Paired Associate Learning - Cogstate	Implicit	Porffy and colleagues [65]	1
Pattern recognition	Continuous Paired Associate Learning - Cogstate	Implicit	Porffy and colleagues [65]	1
Perceptual Motor	None specifically reported	N/A	Tan and colleagues [66]	1
Social Cognition	None specifically reported	N/A	Tan and colleagues [66]	1
Learning and	None specifically reported	N/A	Tan and colleagues	1

memory			[66]	
Language	None specifically reported	N/A	Tan and colleagues [66]	1
Calculation	None specifically reported	N/A	Tsai and colleagues [67]	1

^aBanville and colleagues [61]: The VR task was predominantly a sorting task for executive function assessment. The comparative assessments that validated this assessment have been detailed under 'executive function' broadly as the paper did not specify which components of the VR task the comparative tasks aimed to validate.

^bKourtesis et al [76]: explicitly break episodic memory down into immediate and delayed recognition. However, we have gathered these two constructs under episodic memory.

^cMarín-Morales and colleagues [70]: The VR task involves 42 VR mini-games which assess various cognitive constructs. Four mini-games and their target constructs are documented and included in this table, however the comparative assessments were not provided, and an extensive list of all 42 mini-games not provided.

^dRobitaille and colleagues [74]: Used a VR paradigm with avatars to trial a dual task walking protocol Williams and colleagues [55]: Replicated the WCST and multitasking task, but did not explicitly state the cognitive constructs the VR task was assessing. N/A = Not applicable.

Task index: D-KEFS = Delis-Kaplan Executive Function System [77]; TMT A/B/AB = Trail Making Test A or B or both AB sets; ST = Stroop Task; SET = Six Element Test; HTT = Hanoi Tower Test; ZMT = Zoo Map Test; OTS = One-touch stockings of Cambridge mean choices to correct (CANTAB; Cambridge Cognition Ltd.); VFT = Verbal Fluency Test (S and D) (VFT); DOT = Dot Probe Task; GNG = Go No-Go Task; CPT = Continuous performance Test [78]; CW-IT, D-KEFS = The Color Word Interference Test from the D-KEFS; BIS = Barratt Impulsiveness Scale [79]; WCST = Wisconsin Card Sorting Test; LNS, WAIS-III = Letter-Number Sequencing from the WAIS-III; SWM, CANTAB = Spatial working memory from CANTAB; TOL-DX = Tower of London-Drexler test; TOL = Tower of London; KST, BADS = The Key Search task from the BADS [80]; RAVLT = Rey Auditory Verbal Learning Test; CAMPT = Cambridge Prospective Memory Test [81] CVLT-II = California Verbal Learning Test - Second Edition [82]; RMBT-III = Rivermead Behavioral Memory Test-III [83] RVP, CANTAB = Rapid Visual Processing (RVP) from CANTAB; RBANS-DS = RBANS Digit span subtest; CTT-A/B/AB = Color Trails Test A or B or both AB sets [84]; RBANS-CT = RBANS coding test.

Validity and reliability

Tables 1-5 show details of the current validated comparator task against the novel VR task if it was explicitly provided by authors. Where 'None specifically reported' is stated, the authors of each study did not identify, or denote, a direct comparator. All but two of the studies [72,73] set out to assess multiple constructs. In some cases, the sub-constructs that were assessed were individually validated against existing, validated tasks. In other cases, a suite of existing, validated tasks were included in the analysis for correlation against a variety of sub-constructs being assessed by the VR battery. In these cases, there was no validation at the construct level identified a-priori. In three of the 19 studies, there was no reported validation of the VR paradigm.

Notably, only one study utilised real-life validation criteria besides construct-driven tests to present a validation of their VR paradigm. Specifically, Miskowiak and colleagues [63] functionally assessed

participants using the Functioning Assessment Short Test (FAST) and the Brief University of California, San Diego Performance-based Skills Assessment (UPSA-B). Participants' scores on these assessments were correlated to their performance on the test domains of the VR paradigm called CAVIR. The authors identified that participants' performance on the FAST was negatively associated ($-.17 < r < -.30$, no exact or relative P -values reported) to CAVIR test domains, such as processing speed and working memory, whereas participants' performance on the UPSA-B was positively associated to CAVIR test domains Working memory ($r = .40$, P -value not exactly or relatively reported) and cognition composite ($r(68) = .44$, $P < .001$). Moreover, the authors noted that lower global scores on traditional (i.e., construct-led) neuropsychological tests were negatively associated with FAST scores ($r(121) = -.45$, $P < .001$) and positively associated with UPSA-B scores ($r(68) = .52$, $P < .001$), highlighting the association that lower CAVIR scores were associated with more functional disability indicative from the functional and traditional assessment tools.

The reliability of the VR paradigm was only assessed in one study. This was done by Kourtesis and colleagues [76], who reported good internal reliability (Cronbach's $\alpha = 0.79$) of their VR-EAL paradigm. However, this global internal consistency report does not provide a reliability estimate of the unique cognitive functions targeted in their VR EAL paradigm. Nonetheless, none of the studies reviewed included a test-retest analysis to highlight the reliability of their VR paradigm.

Evaluation of user experience, cybersickness, immersion and engagement

An overview of the measures used to evaluate participants' experiences and wellbeing can be found in Multimedia Appendix 1. Of the 19 studies, five included user experience assessments. To measure participants' virtual presence, experience and well-being, studies administered either the iGroup Presence Questionnaire (IPQ [61]), the Presence Questionnaire (PQ [63,71,74]) or the Slater-Usold-Steed questionnaire (SUSQ [74]). To measure participants' discomfort, studies used the Simulator Sickness Questionnaire (SSQ [61,71,74]) or an adaption of the SSQ, the VR Simulation Sickness Questionnaire (VRSSQ [63]). To evaluate the usability of the virtual environment, studies used the System Usability Scale (SUS [71]). To measure participants' virtual experience and comfort, two studies employed the Virtual Reality Neuroscience Questionnaire (VRNQ [76]).

Few studies investigated whether system usability, virtual presence, or cybersickness affected participants' task performance. For example, Porffy and colleagues [65] measured participants' technical familiarity and found that participants' technical familiarity explained between 10 to 42% of the variability in participants performance on the VStore outcomes Recall, Find, and Select. Conversely put, participants' technical familiarity appeared to influence their performance in the VStore. Kourtesis and colleagues [76] employed questionnaires to evaluate the quality of the VR paradigm, participants' gaming experience and the realism (verisimilitude) and pleasantness of the VR paradigm. The authors identified no relationship between VR experience, gaming experience, and performance on the VR-EAL tasks.

Some studies reported on cybersickness, presence, or usability scores but did not report an analysis studying the relationship between task performance and measures evaluating the VR paradigm. For example, Banville and colleagues [61] recorded participants' sickness and virtual presence, but did not report any test evaluating whether sickness or presence affected task performance. Similarly, Voinescu and colleagues [71] obtained system usability ratings from participants, however no test was reported wherein the effect of usability on task performance was assessed. Lastly, Chicchi Giglioli and colleagues [68] recorded participants' use of technology but did not report an analysis between technology use and task performance.

Finally, some studies evaluated participants' experiences post-hoc although it was not disclosed

whether any validated scales were used. For example, Davison and colleagues [62] measured participants' enjoyment of the VR tasks and their preference for either the VR tasks or the pencil-and-paper task. The authors found that younger participants, rather than older ones, preferred VR tasks over pencil-and-paper tasks. Additionally, 11 out of 40 participants reported having experienced a mild degree of motion sickness. Seven studies, however, did not disclose any information about user experiences [53,62,64,68,72,73].

Discussion

The purpose of this review was to investigate the development and validation of VR assessment tools for EF. Specifically, we examined the components of EF that are assessed with VR, their validation processes, and whether immersion and cybersickness assessments were employed. While research in this domain is proliferating, the results from this review suggest that the process of development and validation varies considerably between studies.

Components of EF assessed by VR paradigms

The terminology used by the studies to describe EF constructs was inconsistent. For example, the most popular construct set assessed by VR comprised the inhibition processes. 'Inhibitory control' encompasses the inhibition of goal-irrelevant stimuli, cognitions, and behavioural responses [6,85]. Two of the key components of inhibitory control are response inhibition and attentional inhibition [86]. Response inhibition was also termed "inhibition control", "prepotent response inhibition", and "motor inhibition", while attentional inhibition was also termed "control of interference", "interference control" and "external interference control". Although these terms are used within the literature [86], the readability and synthesis of the literature would be improved by agreement on a particular term for the same construct. In the same way, several studies examined "EF" broadly, without specifically detailing its components. In these studies, the EF was validated with different measurement tools, suggesting that across studies, EF was defined and used differently in each VR paradigm. Since the constructs that these paradigms aim to assess are not explicitly detailed, this poses a risk of hampering researchers wishing to build upon previous findings.

Further, there was a broad range of constructs that are not commonly considered as EF domains but were reported as components of EF, making it difficult for research to replicate findings of undefined target constructs. For example, some studies reported on verbal learning [63], associate learning pattern recognition [65], perceptual motor, social cognition, language [66] and calculation [67]. While many of these components rely upon EF domains, or underpin those domains, they exist at various levels of abstraction. Thus, while the reviewed studies investigate components at different levels and use different languages, it is possible that they overlap. For example, 'organization' may be an umbrella term for a range of EF domains, each of which uses different terminology for the same concept, such as 'cognitive flexibility', 'flexible updating' and 'working memory'. While 'organization' is not measured as a higher-order version of the sub-components, it is difficult for the research that has examined cognitive flexibility and working memory to be extended. Thus, two studies assessing the same construct are not able to build upon each other's progress.

Recommendation

- Establish a coherent and consistent framework for EF terminology

The Research Domain Criteria Framework (RDoC) developed by the National Institute of Mental Health (NIMH) could serve as a framework to address the recommendation. RDoC was originally created to consolidate the research conducted in various fields of mental health [86]. The framework categorises cognition into six domains and encourages investigation of these domains via different classes of variables, such as behavioral, physiological, and self-report data. This framework encourages a common language and organises findings in such a way that researchers can identify

gaps or discrepancies in the literature and contribute to the ongoing development of the field. This framework indicates the potential benefits of employing a common language for research, although it is not necessarily the only option for this field. Alternatively, researchers could engage in a Delphi study to generate expert-informed consensus of the key constructs of EF that merit investigation in VR paradigms (e.g., see Yücel and colleagues [87], for a Delphi study on neuropsychological assessment for addiction). Nonetheless, the emerging area of VR development for neuropsychological assessments would benefit from using the RDoC framework to coordinate the research process.

Validation of VR for EF

Overall, there was limited reporting on the constructs that were assessed by VR paradigms and the associated validation outcome measures. In some studies, there was inadequate reporting of the constructs the EF paradigm was intended to assess. In others, the same construct was assessed using a variety of traditional tasks. Further, some VR paradigms were intended to replicate real-life, yet were validated against traditional tasks, none of which assess ecological validity. In some studies, the correlations between the VR paradigm and the traditional tasks were incomplete. Finally, sample sizes varied considerably between studies, additionally impacting the evaluation of their psychometrics. These points are expanded upon below.

Several studies examined EF as a broad category, and then validated the paradigm against a variety of traditional tasks, with limited (or no) reporting of which aspect of the VR paradigm the traditional task was intended to validate. For example, some studies [61–63] included traditional tasks to validate their VR paradigm designed to give a broad assessment of EF. However, there was no detail provided about which outcome measure from the VR paradigm was intended to reflect which component of EF. Traditional tasks, which often target one construct, were then correlated against seemingly all outcomes of the VR paradigm outcome measures. While this practice may be beneficial during the exploratory phase of VR paradigm development, failure to correct for multiple comparisons may provide misleading results whereby a correlation is found between two constructs incidentally. Conversely, some traditional tasks assess multiple constructs, which poses a slightly different challenge. For example, if the VR paradigm was broadly assessing EF, but was validated against the Stroop task, it was then unclear whether the VR paradigm was aiming to assess processing speed, attention, inhibitory control, or interference control since the Stroop task could be used to measure all four. Similarly, when these studies used multiple traditional assessments, the reader was expected to presume the target constructs of the EF paradigm, since this wasn't clearly outlined. Poorly defined target constructs, and failure to specify which traditional task is validating which aspect of the VR task produces a literature that is difficult to interpret. Moreover, this general lack of clarity means that future researchers are more likely to invent a new paradigm rather than adopt or extend existing paradigms, creating inefficiency and hampering progress in the field.

Other studies sought to validate their target constructs in the VR paradigm with various standardised tasks. For example, the study by Chicchi Giglioli and colleagues [69] examined attention and inhibition control using the Dot Probe Task, Go/No-Go and Stroop. However, Voinescu and colleagues [71] examined inhibition using the Continuous Performance Test paradigm. Additionally, Marín-Morales and colleagues [70] assessed inhibition using one minigame of their VR paradigm. However, they neither provided details of a specific comparator task for validation purposes nor reported the statistical outcomes. Further, the Dot Probe Task, which is typically used to assess selective attention [88], was used to assess inhibition although its own psychometric properties have been the subject of controversy [89,90]. Although several traditional tasks purport to measure the

same construct (i.e., there is not one task for one construct), the lack of consistency between studies makes it difficult to compare VR platforms. Furthermore, the traditional comparator task used to validate the VR paradigm needs to have sound psychometric properties in its own right to assess the respective construct; when two tasks are being compared to one another, it is unclear which task may be responsible for discrepancies in the outcome [91]. These points are especially pertinent for studies that rely solely on traditional measures to validate tasks in the absence of other validation techniques.

While it is promising to see that VR paradigms are being used for ecologically valid assessments, their validation remains a challenge. For traditional tasks, we assume that a single construct can be assessed by a behavioural task, and performance on that task is linear with the cognitive construct. For a 'function-led' VR task, there is a behavioural task that simulates real-world functioning, which is thought to deteriorate in an EF-declining population. This VR task is not a direct assessment of a target construct; it is a test of a real-world function, such as parking a car. To test convergent validity, the individual would have to park a car in real-life and have their performance assessed similar to the VR task and compared. However, when we use traditional measures to validate the 'function-led' VR measures, we assume that EF can be reliably measured, and the function-led VR task (e.g., parking a car) requires the same EF. Thus, those who perform poorly on a traditional EF task are also expected to perform poorly on real-life tasks requiring EF. Critically, if our results do not show this relationship, it could be that the traditional task is a poor test of EF, the function-led assessment is a poor test of EF, or the EF at hand is not related to the functional task (e.g., parking a car).

These assumptions place substantial weight upon the selection of the traditional task for validating the VR paradigm for predictive validity. Davison and colleagues [62] assessed EF using the Stroop and the Trail Making Task (TMT). They broadly hypothesized that there would be correlations between the traditional measures and the VR paradigm which contained tasks that replicate real-life, such as car parking, arranging seating, and locating items. In the reported results, the Stroop and TMT were correlated with all outcome measures of the VR paradigm. For example, performance on Stroop colour-word related to performance on the second parking simulator task, the number of levels completed on the parking simulator task, and the time taken to place the blue chair in the seating arrangement task. If the Stroop and TMT are not sufficient validators of the functional task, this may generate misleading results about the integrity of the VR paradigm and its ability to sensitively measure EF. Thus, convergent validity of VR tasks would be better assessed by real-life performance on the same task, such as actually parking in a controlled environment. While this may seem to be a resource burden to validation, it could provide integral merit to using the paradigm as a proxy to the real-life task thereafter. Alternative options are to assess convergent validity with other forms of real-life functioning (e.g., self-care, residence, transportation, employment) and diagnostic trajectory [49]. Moreover, predictive validators should be carefully chosen to ensure that their target construct aligns with that thought to be required for the function-led assessment.

Nonetheless, for novel task validation, transparent reporting of all results is crucial for advancing future research. Several studies included in this review (e.g., [61,62,68,69]) reported only statistically significant correlations, leaving unanswered questions due to the omission of non-significant results. For instance, Chicchi Giglioli and colleagues [69] sought to evaluate inhibition control, using the Go/No-go and Stroop asks for validation (both are common tasks for assessing inhibition) and the dot probe task, yet did not report all correlational data in their results table. Such omissions hinder the comprehensive use or meta-analytic application of these findings. Conversely, Chicchi Giglioli and colleagues [68] provided a detailed comparison between each validation task and its corresponding VR task, including the constructs assessed. However, only significant correlations were reported, some of which were between tasks intended to assess disparate constructs, like the correlations between the Wisconsin Sorting Card Task (assessing Cognitive shifting) and VR tasks (measuring Attention and Inhibition Control). While these findings may indicate overlapping constructs in VR tasks, the absence of multiple comparison correction and a

detailed post-hoc analysis of these correlations limits the interpretability and applicability of these results.

Finally, it is worth noting that there was significant variation in sample sizes across the studies reviewed. While it is often accepted that pilot studies or preliminary studies have small sample sizes that often result in underpowered analysis, the utility of the VR paradigms is dependent on sound psychometric properties which require adequate sample sizes and statistical power. As detailed in Multimedia Appendix 1, sample sizes varied from 12 (6 per group) [74] to 103 (divided into two groups) [53]. While the definition of a 'sufficient' sample size may vary between studies and analyses, several of the included VR paradigms would likely require additional validation studies to provide confidence in their psychometric properties.

Recommendation

- Studies should explicitly detail how their VR paradigms are being validated. If a paradigm has multiple components, it is essential to state how each one is being validated. A good example is the Kourtesis and colleagues [76] paper in this review.
- If studies are aiming to validate a VR paradigm for a specific EF construct, they should identify a-priori the precise outcome measures of the VR paradigm that are hypothesised to tap into various EF constructs (e.g., time to completion, number of errors), and then validate them against the appropriate traditional task that also reliability assesses that EF construct.
- Where appropriate, the VR paradigm's real-world task should be validated against both traditional task measures and ecologically valid measures. Ecologically valid measures may include: carer reports, observation assessments and activities of daily living assessments.
- Multiple modes of validation should be employed, including measures that provide predictive power [49], and both carer reports of daily functioning as well as biosensor data should be considered.
- Studies should report all outcomes of validation data (even those in supplementary materials) to ensure transparency of the tools' properties. A concerted effort to increase explicit and transparent reporting would greatly benefit the field.
- In order to validate the VR paradigm, the psychometric properties of the traditional task must be appropriate.
- Studies aiming to evaluate the psychometric properties of their VR paradigm should ensure that they have adequate sample size for powered analysis.

Cybersickness

While VR offers several key advantages over traditional tasks, these systems can also produce adverse effects such as cybersickness. In our review, only four of the 19 studies included an assessment of cybersickness. This is concerning, since cybersickness presents a substantial confound for valid VR assessment; cybersickness has been shown to negatively affect task performance [92,93]. Given that the assessment of EF involves ascertaining a participant's cognitive abilities, the recording of cybersickness is key to ensuring that common side-effects, such as dizziness and vertigo, do not affect the participants' ability to perform at their best on the tasks. Without formal evaluation, the degree to which the participant's experience is altered is unclear. Furthermore, it is unknown at this stage whether cybersickness symptoms affect various client populations differently. For example, it is possible that while a healthy individual may be able to continue assessment with

minor vertigo, a cognitively impaired individual may be more affected, resulting in severely impacted cognitive outcomes. Thus, caution should be exercised when using VR paradigms to ensure the potential benefits of engagement and ecological validity are not realized at the cost of the potential negative effects of cybersickness.

Recommendation

- Future studies include usability data in the form of cybersickness measurements.
- Correlations between cybersickness and participants' task performance could be included as supplementary material that should be accessible to readers, enabling them to better understand how the VR battery is performing.
- Even where a paradigm has already assessed cybersickness, we encourage future researchers using the same paradigm to conduct their own cybersickness assessment. This is because it is still unclear whether cybersickness will have different effects on various populations.
- Clinical researchers and engineers should continue to investigate and report on techniques and technologies that reduce the incidence or severity of cybersickness.

Table 6 provides an overview of the recommendations from this review.

Table 6. Recommendations for future research and practice using VR HMD-based paradigms for EF assessment

Recommendation	Description
Validate against multiple forms	Examples: carer reports, observation assessments, ecological momentary assessment, activities of daily living, physiological sensors and in-vivo studies. Consider longitudinal tracking of participants to establish predictive utility to initially validate the novel paradigm.
Report a-priori how each assessment in the VR paradigm is being validated	If there are multiple components to one paradigm, state how each element is being validated (a good example is the Kourtesis and colleagues [76] paper in this review). For example, "Task 2a aims to assess inhibitory control and is validated against the traditional stop signal task and go/no-go task".
Report all validation data	Report correlations of all aspects of a task that were identified a-priori as validating the VR paradigm. In extending the example above, show all relevant metrics from Task 2a such as errors, proportion of successful stops, reaction time, and stop signal reaction time etc. against the relevant metrics of both Stop signal task and Go/No-Go tasks.
Include user-experience assessment	Conduct assessment of: immersion, cybersickness, usability and engagement.
Use a common framework for defining target constructs	The Research Domain Criteria is one option for a framework that can be applied to ensure that terminology used by the field is consistent.
Consider adding biosensors	These provide additional objective data which may inform the VR-based EF assessment.

Limitations

We searched for articles that used the terms "executive functioning", "higher order cognition" and "functional assessment" to capture tasks that were aiming to broadly assess facets of EF. This search

strategy may have missed studies that examined a key construct of EF, but did not specifically use the aforementioned terms (for example, used VR to assess inhibitory control alone). Additionally, we did not contact authors of the papers included in these studies for further information, however one of the key outcomes for this review is the amount of information contained in the manuscripts for future studies to extend upon.

Future directions

The authors posit that the integration of biosensors into a VR system has huge potential. Biosensors, such as pupillometry, eye gaze, electroencephalography (EEG), but also language and grammatical-characteristics data, can be temporally linked with the events occurring in the VR task. For example, pupillometry can offer insights into brain injury prognosis [94], and differentiate between Alzheimer's and healthy controls [95]. Eye-tracking during reading aids Alzheimer's identification [96], and linguistic attributes (e.g., formation and fluency of sentences, syntax and grammar) distinguish Alzheimer's from MCI patients [97]. The combination of these biosensor metrics and real-time function-led VR performance could increase the sensitivity of tests enabling the detection of subtle changes, such as between MCI and subjective memory complaints [98]. However, currently, biosignals are rarely evaluated alongside emerging VR paradigms for EF assessment. None of the reviewed studies used biosensors, leaving an untapped potential for VR paradigms to be front-line neuropsychological assessments.

Biosensors could also assist in modulating the cognitive load experienced by the participant. Cognitive load is the cumulative working memory resources that an individual requires for a given task [99]. Similar to the gaming industry, VR paradigms could be adaptive and performance-driven so that the level of challenge adjusts according to real-time individual responses [100,101]. Modulating the cognitive load adjusts the challenge of a task and enables all participants to encounter similar levels of perceived difficulty for their respective abilities. EEG, pupillometry, and cardiovascular measures are also sensitive to cognitive load capacities [99].

An additional advantage of VR is its ability to facilitate the assessment of spatial navigation. Spatial navigation is a component of cognitive functioning that can be a key factor in detecting early stages of neurodegenerative disease. However, it cannot be assessed adequately by means of many traditional assessments. While it is acknowledged that spatial navigation is not a component of executive functioning, the authors of this paper consider it a generally under-examined construct when assessing cognition and general function. For example, spatial navigation is a cognitive marker used to detect early attention deficit [102,103], and offers additional relevant information beyond the traditional neuropsychological tests [103]. The environment could also be systematically manipulated to match the needs of the assessment [104] and tailored to specific populations. However, typically, spatial navigation is assessed using a real-space human analogue of the Morris Water Maze test, which can be difficult to implement under standardised conditions. Computerised versions have been adapted with findings comparable to those of tests done in real-space [105], suggesting promise for translating this style of assessment into VR.

Conclusion

VR paradigms assessing EF have great potential to improve upon traditional tests. However, despite undeniable novelty and potential, their methodological and psychometric properties must be addressed during their development to ensure their validity and reliability. Although there is no shortage of research in this area, the lack of standardized protocols to validate VR-based neuropsychological assessments hinder the progress of this field of research and practice. It is hoped that this paper is the beginning of a larger movement towards systematising the development and validation of these paradigms.

Acknowledgements

Funding information: LK receives funding support from the Institute for Intelligent Systems Research and Innovation. This funding source had no role in the data analysis, presentation, or interpretation and write-up of the data. MY receives funding from: government funding bodies such as the NHMRC, Australian Research Council (ARC), Australian Defence Science and Technology (DST), the Department of Industry, Innovation and Science (DIIS), the National Institutes of Health (NIH, USA); philanthropic donations from the David Winston Turner Endowment Fund, Wilson Foundation; sponsored Investigator-Initiated trials including Incannex Healthcare Ltd; and payments in relation to court-, expert witness-, and/or expert review-reports. These funding sources had no role in the data analysis, presentation, or interpretation and write-up of the data. Murat Yücel also sits on the Advisory Boards of: Centre of The Urban Mental Health, University of Amsterdam; Monash Biomedical Imaging Centre; and Enosis Therapeutics. The authors would additionally like to acknowledge Scribo Proofreading and Editing for providing their services on this manuscript.

Author Contribution Statement: RK, KR and LK made significant and equal contributions to the concept and design of the systematic review. RK and KR independently screened abstracts and full texts; RK and LK completed the data extraction process using Excel and manually reviewed each manuscript. RK, KR and LK contributed to the manuscript and LA, MY, DM provided substantial critical feedback on the manuscript.

Conflicts of interest

None declared

Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

References

1. Anderson P. Assessment and Development of Executive Function (EF) During Childhood. *Child Neuropsychology* 2002;8:71–82. <https://doi.org/10.1076/chin.8.2.71.8724>.
2. Lehto JE, Juujärvi P, Kooistra L, Pulkkinen L. Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology* 2003;21:59–80. <https://doi.org/10.1348/026151003321164627>.
3. Manchester D, Priestley N, Jackson H. The assessment of executive functions: coming out of the office. *Brain Injury* 2004;18:1067–81. <https://doi.org/10.1080/02699050410001672387>.
4. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The Unity and Diversity of Executive Functions and Their Contributions to Complex “Frontal Lobe” Tasks: A Latent Variable Analysis. *Cognitive Psychology* 2000;41:49–100. <https://doi.org/10.1006/cogp.1999.0734>.
5. Suchy Y. Executive Functioning: Overview, Assessment, and Research Issues for Non-Neuropsychologists. *Annals of Behavioral Medicine* 2009;37:106–16. <https://doi.org/10.1007/s12160-009-9097-4>.
6. Diamond A. Executive Functions. *Annual Review of Psychology* 2013;64:135–68. <https://doi.org/10.1146/annurev-psych-113011-143750>.
7. Collins A, Koechlin E. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making. *PLOS Biology* 2012;10:e1001293. <https://doi.org/10.1371/journal.pbio.1001293>.
8. Lunt L, Bramham J, Morris RG, Bullock PR, Selway RP, Xenitidis K, et al. Prefrontal cortex dysfunction and ‘Jumping to Conclusions’: Bias or deficit? *Journal of Neuropsychology* 2012;6:65–78. <https://doi.org/10.1111/j.1748-6653.2011.02005.x>.
9. Elliott R. Executive functions and their disorders: Imaging in clinical neuroscience. *British Medical Bulletin* 2003;65:49–59. <https://doi.org/10.1093/bmb/65.1.49>.
10. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology* 2015;6:328. <https://doi.org/10.3389/fpsyg.2015.00328>.
11. Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review. *Psychological Bulletin* 2013;139:81–132. <https://doi.org/10.1037/a0028727>.
12. Demetriou EA, Lampit A, Quintana DS, Naismith SL, Song YJC, Pye JE, et al. Autism spectrum disorders: a meta-analysis of executive function. *Mol Psychiatry* 2018;23:1198–204. <https://doi.org/10.1038/mp.2017.75>.
13. Chan RCK, Shum D, Touloupoulou T, Chen EYH. Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology* 2008;23:201–16. <https://doi.org/10.1016/j.acn.2007.08.010>.
14. Burgess PW, Alderman N, Forbes C, Costello A, M-A.Coates L, Dawson DR, et al. The case for the development and use of “ecologically valid” measures of executive function in experimental and clinical neuropsychology. *J Int Neuropsychol Soc* 2006;12:194–209. <https://doi.org/10.1017/S1355617706060310>.
15. Burgess PW, Alderman N, Evans J, Emslie H, Wilson BA. The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society* 1998;4:547–58. <https://doi.org/10.1017/S1355617798466037>.
16. Sbordone RJ. Ecological validity: Some critical issues for the neuropsychologist. *Ecological validity of neuropsychological testing*, Delray Beach, FL, England: Gr Press/St Lucie Press, Inc; 1996, p. 15–41.
17. Kvavilashvili L, Ellis JA. Ecological validity and the real-life/laboratory controversy in memory research: A critical and historical review. *History and Philosophy of Psychology*

- 2004;6:59–80. ISSN 1469-283X
18. Spooner D, Pachana N. Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology* 2006;21:327–37. <https://doi.org/10.1016/j.acn.2006.04.004>.
 19. Chaytor N, Schmitter-Edgecombe M, Burr R. Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology* 2006;21:217–27. <https://doi.org/10.1016/j.acn.2005.12.002>.
 20. Gärtner A, Strobel A. Individual Differences in Inhibitory Control: A latent Variable Analysis. *J Cogn* 2021;4:17. <https://doi.org/10.5334/joc.150>.
 21. Choi SY, Lee JS, Oh YJ. Cut-off point for the trail making test to predict unsafe driving after stroke. *Journal of Physical Therapy Science* 2016;28:2110–3. <https://doi.org/10.1589/jpts.28.2110>.
 22. Papandonatos GD, Ott BR, Davis JD, Barco PP, Carr DB. The Clinical Utility of the Trail Making Tests As Predictors of Driving Performance in Older Adults. *J Am Geriatr Soc* 2015;63:2358–64. <https://doi.org/10.1111/jgs.13776>.
 23. Dobbs BM, Shergill SS. How effective is the Trail Making Test (Parts A and B) in identifying cognitively impaired drivers? *Age and Ageing* 2013;42:577–81. <https://doi.org/10.1093/ageing/aft073>.
 24. Duncanson H, Hollis AM, O'Connor MG. Errors versus speed on the trail making test: Relevance to driving performance. *Accident Analysis & Prevention* 2018;113:125–30. <https://doi.org/10.1016/j.aap.2018.01.004>.
 25. Vaucher P, Herzig D, Cardoso I, Herzog MH, Mangin P, Favrat B. The trail making test as a screening instrument for driving performance in older drivers; a translational research. *BMC Geriatr* 2014;14:123. <https://doi.org/10.1186/1471-2318-14-123>.
 26. Bublak P, Redel P, Sorg C, Kurz A, Förstl H, Müller HJ, et al. Staged decline of visual processing capacity in mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging* 2011;32:1219–30. <https://doi.org/10.1016/j.neurobiolaging.2009.07.012>.
 27. Buracchio T, Dodge HH, Howieson D, Wasserman D, Kaye J. The Trajectory of Gait Speed Preceding Mild Cognitive Impairment. *Archives of Neurology* 2010;67:980–6. <https://doi.org/10.1001/archneurol.2010.159>.
 28. Allcoat D, Mühlenen A von. Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology* 2018;26. <https://doi.org/10.25304/rlt.v26.2140>.
 29. Huang W, Roscoe RD, Johnson-Glenberg MC, Craig SD. Motivation, engagement, and performance across multiple virtual reality sessions and levels of immersion. *Journal of Computer Assisted Learning* 2021;37:745–58. <https://doi.org/10.1111/jcal.12520>.
 30. Parsons TD, Phillips AS. Virtual reality for psychological assessment in clinical practice. *Practice Innovations* 2016;1:197–217. <https://doi.org/10.1037/pri0000028>.
 31. Shallice T, Burgess P. Deficits in strategy application following frontal lobe damage in man. *Brain: A Journal of Neurology* 1991;114 (Pt 2):727–41. <https://doi.org/10.1093/brain/114.2.727>.
 32. Dawson DR, Anderson ND, Burgess P, Cooper E, Krpan KM, Stuss DT. Further Development of the Multiple Errands Test: Standardized Scoring, Reliability, and Ecological Validity for the Baycrest Version. *Archives of Physical Medicine and Rehabilitation* 2009;90:S41–51. <https://doi.org/10.1016/j.apmr.2009.07.012>.
 33. Maeir A, Krauss S, Katz N. Ecological Validity of the Multiple Errands Test (MET) on Discharge from Neurorehabilitation Hospital. *OTJR: Occupational Therapy Journal of Research* 2011;31:S38–46. <https://doi.org/10.3928/15394492-20101108-07>.
 34. Antoniuk K, Clores J, Jensen D, Nalder E, Rotenberg S, Dawson DR. Developing and Validating a Big-Store Multiple Errands Test. *Frontiers in Psychology* 2019;10.

- <https://doi.org/10.3389/fpsyg.2019.02575>
35. Nalder EJ, Clark AJ, Anderson ND, Dawson DR. Clinicians' perceptions of the clinical utility of the Multiple Errands Test for adults with neurological conditions. *Neuropsychological Rehabilitation* 2017;27:685–706. <https://doi.org/10.1080/09602011.2015.1067628>.
 36. Draschkow D. Remote virtual reality as a tool for increasing external validity. *Nat Rev Psychol* 2022;1:433–4. <https://doi.org/10.1038/s44159-022-00082-8>.
 37. Albani G, Riva G. Virtual multiple errands test: reliability, usability and possible applications. *Annual Review of Cybertherapy and Telemedicine* 2013: Positive Technology and Health Engagement for Healthy Living and Active Ageing 2013;191:38. <https://doi.org/10.3233/978-1-61499-282-0-38>.
 38. Cipresso P, Albani G, Serino S, Pedroli E, Pallavicini F, Mauro A, et al. Virtual multiple errands test (VMET): a virtual reality-based tool to detect early executive functions deficit in Parkinson's disease. *Frontiers in Behavioral Neuroscience* 2014;8. PMID: 25538578
 39. Riva G, Gaggioli A, Grassi A, Raspelli S, Cipresso P, Pallavicini F, et al. NeuroVR 2—a free virtual reality platform for the assessment and treatment in behavioral health care. *Medicine meets virtual reality* 18, IOS Press; 2011, p. 493–5. <https://doi.org/10.3233/978-1-60750-706-2-493>.
 40. Mount NJ, Chambers C, Weaver D, Priestnall G. Learner immersion engagement in the 3D virtual world: principles emerging from the DELVE project. *Innovation in Teaching and Learning in Information and Computer Sciences* 2009;8:40–55. <https://doi.org/10.11120/ital.2009.08030040>.
 41. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of Cognitive Assessment and Cognitive Training: A Systematic Review of Applications and Efficacy. *JMIR Serious Games* 2016;4:e5888. <https://doi.org/10.2196/games.5888>.
 42. Krath J, Schürmann L, von Korflesch HFO. Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior* 2021;125:106963. <https://doi.org/10.1016/j.chb.2021.106963>.
 43. Vermeir JF, White MJ, Johnson D, Crombez G, Ryckeghem DMLV. The Effects of Gamification on Computerized Cognitive Training: Systematic Review and Meta-Analysis. *JMIR Serious Games* 2020;8:e18644. <https://doi.org/10.2196/18644>.
 44. Llobera J, Jacquat V, Calabrese C, Charbonnier C. Playing the mirror game in virtual reality with an autonomous character. *Sci Rep* 2022;12:21329. <https://doi.org/10.1038/s41598-022-25197-z>.
 45. Mahtab EAF, Egorova AD. Current and future applications of virtual reality technology for cardiac interventions. *Nat Rev Cardiol* 2022;19:779–80. <https://doi.org/10.1038/s41569-022-00789-4>.
 46. Li G, Anguera JA, Javed SV, Khan MA, Wang G, Gazzaley A. Enhanced Attention Using Head-mounted Virtual Reality. *Journal of Cognitive Neuroscience* 2020;32:1438–54. https://doi.org/10.1162/jocn_a_01560.
 47. Nalivaiko E, Davis SL, Blackmore KL, Vakulin A, Nesbitt KV. Cybersickness provoked by head-mounted display affects cutaneous vascular tone, heart rate and reaction time. *Physiology & Behavior* 2015;151:583–90. <https://doi.org/10.1016/j.physbeh.2015.08.043>.
 48. Sepich NC, Jasper A, Fieffer S, Gilbert SB, Dorneich MC, Kelly JW. The impact of task workload on cybersickness. *Frontiers in Virtual Reality* 2022;3. <https://doi.org/10.3389/frvir.2022.943409>
 49. Neşuţ A, Matu S-A, Sava FA, David D. Virtual reality measures in neuropsychological assessment: a meta-analytic review. *The Clinical Neuropsychologist* 2016;30:165–84. <https://doi.org/10.1080/13854046.2016.1144793>.

50. Borgnis F, Baglio F, Pedroli E, Rossetto F, Uccellatore L, Oliveira JAG, et al. Available Virtual Reality-Based Tools for Executive Functions: A Systematic Review. *Front Psychol* 2022;13:833136. <https://doi.org/10.3389/fpsyg.2022.833136>.
51. Kim E, Han J, Choi H, Prié Y, Vigier T, Bulteau S, et al. Examining the Academic Trends in Neuropsychological Tests for Executive Functions Using Virtual Reality: Systematic Literature Review. *JMIR Serious Games* 2021;9:e30249. <https://doi.org/10.2196/30249>.
52. Urbina S. *Essentials of Psychological Testing*. John Wiley & Sons; 2014. ISBN: 9781118873090
53. Parsons TD, McMahan T. An initial validation of the Virtual Environment Grocery Store. *Journal of Neuroscience Methods* 2017;291:13–9. <https://doi.org/10.1016/j.jneumeth.2017.07.027>.
54. Wilf M, Ben Yair N, Wright WG, Plotnik M. The trail less traveled: Analytical approach for creating shortened versions for virtual reality-based color trails test. *Applied Neuropsychology: Adult* 2022;0:1–10. <https://doi.org/10.1080/23279095.2022.2065204>.
55. Williams RM, Alikhademi K, Gilbert JE. Design of a toolkit for real-time executive function assessment in custom-made virtual experiences and interventions. *International Journal of Human-Computer Studies* 2022;158:102734. <https://doi.org/10.1016/j.ijhcs.2021.102734>.
56. Baggetta P, Alexander PA. Conceptualization and Operationalization of Executive Function. *Mind, Brain, and Education* 2016;10:10–33. <https://doi.org/10.1111/mbe.12100>.
57. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group*. The Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine* 2009;151:264–9. PMID: PMC2707599
58. Zhu J, Liu W. A tale of two databases: the use of Web of Science and Scopus in academic papers. *Scientometrics* 2020;123:321–35. <https://doi.org/10.1007/s11192-020-03387-8>.
59. Mongeon P, Paul-Hus A. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics* 2016;106:213–28. <https://doi.org/10.1007/s11192-015-1765-5>.
60. Babineau J. Product Review: Covidence (Systematic Review Software). *Journal of the Canadian Health Libraries Association / Journal de l'Association Des Bibliothèques de La Santé Du Canada* 2014;35:68–71. <https://doi.org/10.5596/c14-016>.
61. Banville F, Lussier C, Massicotte E, Verhulst E, Couture J-F, Allain P, et al. Validation of a Sorting Task Implemented in the Virtual Multitasking Task-2 and Effect of Aging. In: Yamamoto S, Mori H, editors. *Human Interface and the Management of Information. Information in Applications and Services*, Cham: Springer International Publishing; 2018, p. 41–54. https://doi.org/10.1007/978-3-319-92046-7_4.
62. Davison SMC, Deeprose C, Terbeck S. A comparison of immersive virtual reality with traditional neuropsychological measures in the assessment of executive functions. *Acta Neuropsychiatrica* 2018;30:79–89. <https://doi.org/10.1017/neu.2017.14>.
63. Miskowiak KW, Jespersen AE, Kessing LV, Aggestrup AS, Glenthøj LB, Nordentoft M, et al. Cognition Assessment in Virtual Reality: Validity and feasibility of a novel virtual reality test for real-life cognitive functions in mood disorders and psychosis spectrum disorders. *Journal of Psychiatric Research* 2022;145:182–9. <https://doi.org/10.1016/j.jpsychires.2021.12.002>.
64. Pallavicini F, Pepe A, Minissi ME. Taking Neuropsychological Test to the Next Level: Commercial Virtual Reality Video Games for the Assessment of Executive Functions. In: Antona M, Stephanidis C, editors. *Universal Access in Human-Computer Interaction. Multimodality and Assistive Environments*, Cham: Springer International Publishing; 2019, p. 133–49. https://doi.org/10.1007/978-3-030-23563-5_12.
65. Porffy LA, Mehta MA, Patchitt J, Boussebaa C, Brett J, D'Oliveira T, et al. A Novel Virtual Reality Assessment of Functional Cognition: Validation Study. *J Med Internet Res* 2022;24:e27641. <https://doi.org/10.2196/27641>.
66. Tan NC, Lim JE, Allen JC, Wong WT, Quah JHM, Muthulakshmi P, et al. Age-Related

- Performance in Using a Fully Immersive and Automated Virtual Reality System to Assess Cognitive Function. *Front Psychol* 2022;13:847590. <https://doi.org/10.3389/fpsyg.2022.847590>.
67. Tsai C-F, Chen C-C, Wu EH-K, Chung C-R, Huang C-Y, Tsai P-Y, et al. A Machine-Learning-Based Assessment Method for Early-Stage Neurocognitive Impairment by an Immersive Virtual Supermarket. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2021;29:2124–32. <https://doi.org/10.1109/TNSRE.2021.3118918>.
 68. Chicchi Giglioli IA, de Juan Ripoll C, Parra E, Alcañiz Raya M. Are 3D virtual environments better than 2D interfaces in serious games performance? An explorative study for the assessment of executive functions. *Applied Neuropsychology: Adult* 2021;28:148–57. <https://doi.org/10.1080/23279095.2019.1607735>.
 69. Chicchi Giglioli IA, Pérez Gálvez B, Gil Granados A, Alcañiz Raya M. The Virtual Cooking Task: A Preliminary Comparison Between Neuropsychological and Ecological Virtual Reality Tests to Assess Executive Functions Alterations in Patients Affected by Alcohol Use Disorder. *Cyberpsychology, Behavior, and Social Networking* 2021;24:673–82. <https://doi.org/10.1089/cyber.2020.0560>.
 70. Marín-Morales J, Carrasco-Ribelles LA, Alcañiz M, Giglioli IAC. Applying machine learning to a virtual serious game for neuropsychological assessment. 2021 IEEE Global Engineering Education Conference (EDUCON), 2021, p. 946–9. <https://doi.org/10.1109/EDUCON46332.2021.9454138>.
 71. Voinescu A, Petrini K, Stanton Fraser D, Lazarovicz R-A, Papavă I, Fodor LA, et al. The effectiveness of a virtual reality attention task to predict depression and anxiety in comparison with current clinical measures. *Virtual Reality* 2021. <https://doi.org/10.1007/s10055-021-00520-7>.
 72. Parsons TD, Carlew AR. Bimodal Virtual Reality Stroop for Assessing Distractor Inhibition in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders* 2016;46:1255-.
 73. Parsons TD, Barnett M. Virtual Apartment-Based Stroop for assessing distractor inhibition in healthy aging. *Applied Neuropsychology: Adult* 2019;26:144–54. <https://doi.org/10.1080/23279095.2017.1373281>.
 74. Robitaille N, Jackson PL, Hébert LJ, Mercier C, Bouyer LJ, Fecteau S, et al. A Virtual Reality avatar interaction (VRai) platform to assess residual executive dysfunction in active military personnel with previous mild traumatic brain injury: proof of concept. *Disability and Rehabilitation: Assistive Technology* 2017;12:758–64. <https://doi.org/10.1080/17483107.2016.1229048>.
 75. Kourtesis P, MacPherson SE. An ecologically valid examination of event-based and time-based prospective memory using immersive virtual reality: The influence of attention, memory, and executive function processes on real-world prospective memory. *Neuropsychological Rehabilitation* 2021;0:1–26. <https://doi.org/10.1080/09602011.2021.2008983>.
 76. Kourtesis P, Collina S, Dumas LAA, MacPherson SE. Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An Immersive Virtual Reality Neuropsychological Battery with Enhanced Ecological Validity. *Journal of the International Neuropsychological Society* 2021;27:181–96. <https://doi.org/10.1017/S1355617720000764>.
 77. Delis DC, Kaplan E, Kramer JH. Delis-Kaplan executive function system. *Assessment* 2001. <https://doi.org/10.1037/t15082-000>
 78. Mueller ST, Piper BJ. The psychology experiment building language (PEBL) and PEBL test battery. *Journal of Neuroscience Methods* 2014;222:250–9. PMID: PMC3897935
 79. Barratt ES. Anxiety and impulsiveness related to psychomotor efficiency. *Perceptual and Motor Skills* 1959;9:191–8. <https://doi.org/10.2466/pms.1959.9.3.191>.

80. Wilson B, Alderman N, Burgess P, Emslie H, Evans J. Behavioural assessment of the dysexecutive syndrome (BADS), thames valley test company. Bury St Edmunds, England 1996.
81. Wilson BA. The cambridge prospective memory test: CAMPROMPT. Pearson Assessment; 2005.
82. Delis D, Kramer J, Kaplan E, Ober B. California verbal learning test. The psychological corporation. San Antonio, TX 2000.
83. Wilson B, Greenfield E, Clare L, Baddeley A, Cockburn J, Watson P, et al. Rivermead behavioural memory test-(RBMT-3) london 2008.
84. D'Elia L, Satz P, Uchiyama C, White T. Color trails test: CTT: psychological assessment resources odessa 1996.
85. Nigg JT. Annual Research Review: On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *Journal of Child Psychology and Psychiatry* 2017;58:361–83. <https://doi.org/10.1111/jcpp.12675>.
86. Tiego J, Testa R, Bellgrove MA, Pantelis C, Whittle S. A Hierarchical Model of Inhibitory Control. *Front Psychol* 2018;9:1339. <https://doi.org/10.3389/fpsyg.2018.01339>.
87. Yücel M, Oldenhof E, Ahmed SH, Belin D, Billieux J, Bowden-Jones H, et al. A transdiagnostic dimensional approach towards a neuropsychological assessment for addiction: an international Delphi consensus study. *Addiction* 2019;114:1095–109. <https://doi.org/10.1111/add.14424>.
88. Kappenman ES, Farrens JL, Luck SJ, Proudfit GH. Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. *Frontiers in Psychology* 2014;5:1368. <https://doi.org/10.3389/fpsyg.2014.01368>.
89. Schmukle SC. Unreliability of the dot probe task. *Eur J Pers* 2005;19:595–605. <https://doi.org/10.1002/per.554>.
90. Chapman A, Devue C, Grimshaw GM. Fleeting reliability in the dot-probe task. *Psychological Research* 2019;83:308–20. <https://doi.org/10.1007/s00426-017-0947-6>.
91. Abma IL, Rovers M, van der Wees PJ. Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes. *BMC Research Notes* 2016;9:226. <https://doi.org/10.1186/s13104-016-2034-2>.
92. Pöhlmann K, Li G, McGill M, Markoff R, Brewster S. You Spin Me Right Round, Baby, Right Round: Examining the Impact of Multi-Sensory Self-Motion Cues on Motion Sickness During a VR Reading Task 2023:712. <https://eprints.gla.ac.uk/291787/> (accessed November 21, 2023).
93. Kim YY, Kim HJ, Kim EN, Ko HD, Kim HT. Characteristic changes in the physiological components of cybersickness. *Psychophysiology* 2005;42:616–25. <https://doi.org/10.1111/j.1469-8986.2005.00349.x>.
94. Solari D, Miroz J-P, Oddo M. Opening a Window to the Injured Brain: Non-invasive Neuromonitoring with Quantitative Pupillometry. In: Vincent J-L, editor. *Annual Update in Intensive Care and Emergency Medicine* 2018, Cham: Springer International Publishing; 2018, p. 503–18. https://doi.org/10.1007/978-3-319-73670-9_38.
95. Pavisic IM, Firth NC, Parsons S, Rego DM, Shakespeare TJ, Yong KXX, et al. Eyetracking Metrics in Young Onset Alzheimer's Disease: A Window into Cognitive Visual Functions. *Frontiers in Neurology* 2017;8. <https://doi.org/10.3389/fneur.2017.00377>.
96. Fernández G, Manes F, Politi LE, Orozco D, Schumacher M, Castro L, et al. Patients with Mild Alzheimer's Disease Fail When Using Their Working Memory: Evidence from the Eye Tracking Technique. *Journal of Alzheimer's Disease* 2016;50:827–38. <https://doi.org/10.3233/JAD-150265>.
97. Fraser KC, Meltzer JA, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in

- Narrative Speech. JAD 2015;49:407–22. <https://doi.org/10.3233/JAD-150520>.
98. Iliadou P, Paliokas I, Zygouris S, Lazarou E, Votis K, Tzovaras D, et al. A Comparison of Traditional and Serious Game-Based Digital Markers of Cognition in Older Adults with Mild Cognitive Impairment and Healthy Controls. *Journal of Alzheimer's Disease* 2021;79:1747–59. <https://doi.org/10.3233/JAD-201300>.
 99. Ranchet M, Morgan JC, Akinwuntan AE, Devos H. Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures. *Neuroscience & Biobehavioral Reviews* 2017;80:516–37. <https://doi.org/10.1016/j.neubiorev.2017.07.001>.
 100. Mishra J, Anguera JA, Gazzaley A. Video Games for Neuro-Cognitive Optimization. *Neuron* 2016;90:214–8. <https://doi.org/10.1016/j.neuron.2016.04.010>.
 101. LaConte SM. Decoding fMRI brain states in real-time. *NeuroImage* 2011;56:440–54. <https://doi.org/10.1016/j.neuroimage.2010.06.052>.
 102. Benke T, Karner E, Petermichl S, Prantner V, Kemmler G. Neuropsychological Deficits Associated With Route Learning in Alzheimer Disease, MCI, and Normal Aging. *Alzheimer Disease & Associated Disorders* 2014;28:162. <https://doi.org/10.1097/WAD.0000000000000009>.
 103. Laczó J, Andel R, Nedelska Z, Vyhnalek M, Vlcek K, Crutch S, et al. Exploring the contribution of spatial navigation to cognitive functioning in older adults. *Neurobiology of Aging* 2017;51:67–70. <https://doi.org/10.1016/j.neurobiolaging.2016.12.003>.
 104. Cogné M, Taillade M, N’Kaoua B, Tarruella A, Klinger E, Larrue F, et al. 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* 2017;60:164–76. <https://doi.org/10.1016/j.rehab.2015.12.004>.
 105. Schoenfeld R, Foreman N, Leplow B. Ageing and spatial reversal learning in humans: Findings from a virtual water maze. *Behavioural Brain Research* 2014;270:47–55. <https://doi.org/10.1016/j.bbr.2014.04.036>.

Abbreviations

EF = Executive Function
 VR = Virtual Reality
 WoS = Web of Science
 D-KEFS = Delis-Kaplan Executive Function System (Delis et al., 2001)
 TMT A/B/AB = Trail Making Test A or B or both AB sets
 ST = Stroop Task
 SET = Six Element Test
 HTT = Hanoi Tower Test
 ZMT = Zoo Map Test
 OTS = One-touch stockings of Cambridge mean choices to correct (CANTAB; Cambridge Cognition Ltd.)
 VFT = Verbal Fluency Test (S and D) (VFT)
 DOT = Dot Probe Task; GNG = Go No-Go Task
 CPT = Continuous performance Test (Mueller and Piper, 2014)
 CW-IT, D-KEFS = The Color Word Interference Test from the D-KEFS
 BIS = Barratt Impulsiveness Scale (Barratt, 1959)
 WCST = Wisconsin Card Sorting Test
 LNS, WAIS-III = Letter-Number Sequencing from the WAIS-III
 SWM, CANTAB = Spatial working memory from CANTAB
 TOL-DX = Tower of London-Drexler test
 TOL = Tower of London
 KST, BADS = The Key Search task from the BADS (Wilson et al., 1996)
 RAVLT = Rey Auditory Verbal Learning Test
 CAMPROPT = Cambridge Prospective Memory Test (Wilson et al., 2005)
 CVLT-II = California Verbal Learning Test - Second Edition (Delis et al., 2000)
 RMBT-III = Rivermead Behavioral Memory Test-III (Wilson et al., 2008)
 RVP, CANTAB = Rapid Visual Processing (RVP) from CANTAB
 RBANS-DS = RBANS Digit span subtest
 CTT-A/B/AB = Color Trails Test A or B or both AB sets (D’Elia et al., 1996)

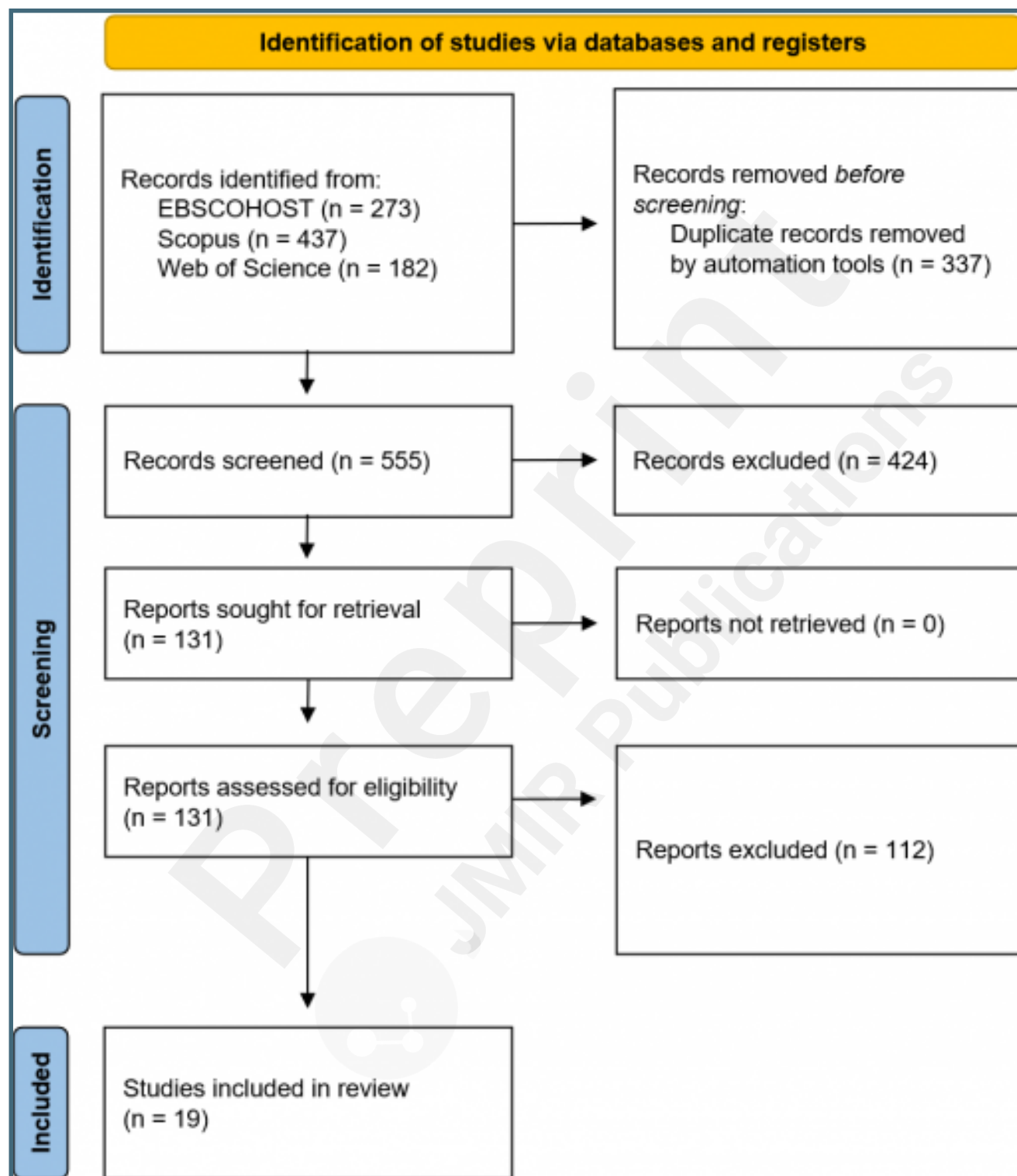
RBANS-CT = RBANS coding test



Supplementary Files

Figures

Systematic review process and results from literature searches in EBSCOHost, Scopus and Web of Science databases.



Multimedia Appendixes

Article details including participant demographics, VR paradigm, VR tasks, measures of user experience and comparative assessments for VR paradigms.

URL: <http://asset.jmir.pub/assets/4933333f409a61d8d0a744f83880c417.docx>



License to publish form (upload after acceptance)s

PRISMA checklist.

URL: <http://asset.jmir.pub/assets/6dd30c953abc1ebdf8b5e832f2739d06.pdf>

PRISMA abstract checklist.

URL: <http://asset.jmir.pub/assets/363d14852d79e627bff808e8468cddf8.pdf>

Search Strategy.

URL: <http://asset.jmir.pub/assets/6dfc3346b64ff44ea16ee712bdbfb5fc.pdf>