Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity

Panagiotis Kourtesis, Simona Collina, Leonidas A.A. Doumas, Sarah E Macpherson

Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity

Panagiotis Kourtesis[a,b,c,d]*, Simona Collina[c,d], Leonidas A.A. Doumas[b], and Sarah E. MacPherson[a,b].

[a] *Human Cognitive Neuroscience, Department of Psychology, University of Edinburgh, Edinburgh, UK;*

[b] *Department of Psychology, University of Edinburgh, Edinburgh, UK;*

[c] *Lab of Experimental Psychology, Suor Orsola Benincasa University of Naples, Naples, Italy;*

[d] *Interdepartmental Centre for Planning and Research "Scienza Nuova", Suor Orsola Benincasa University of Naples, Naples, Italy*

* Panagiotis Kourtesis, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, Scotland, United Kingdom

Email: pkourtes@exseed.ed.ac.uk

Abstract

Objective: The assessment of cognitive functions such as prospective memory, episodic memory, attention, and executive functions benefits from an ecologically valid approach to better understand how performance outcomes generalize to everyday life. Immersive virtual reality (VR) is considered capable of simulating real-life situations to enhance ecological validity. The present study attempted to validate the Virtual Reality Everyday Assessment Lab (VR-EAL), an immersive VR neuropsychological battery, against an extensive paper-and-pencil neuropsychological battery.

Methods: Forty-one participants (21 females) were recruited: 18 gamers and 23 non-gamers who attended both an immersive VR and a paper-and-pencil testing session. Bayesian Pearson correlation analyses were conducted to assess construct and convergent validity of the VR-EAL. Bayesian t-tests were performed to compare VR and paper-and-pencil testing in terms of administration time, similarity to real life tasks (i.e., ecological validity), and pleasantness.

Results: VR-EAL scores were significantly correlated with their equivalent scores on the paper-and-pencil tests. The participants' reports indicated that the VR-EAL tasks were significantly more ecologically valid and pleasant than the paper-and-pencil neuropsychological battery. The VR-EAL battery also had a shorter administration time.

Conclusion: The VR-EAL appears as an effective neuropsychological tool for the assessment of everyday cognitive functions, which has enhanced ecological validity, a highly pleasant testing experience, and does not induce cybersickness.

Keywords: Prospective Memory, Episodic Memory, Attention, Executive Function, Everyday Functioning, Virtual Reality.

Introduction

The ability to perform activities in everyday life is dependent upon cognitive abilities such as attention, episodic memory, executive abilities and prospective memory (Mlinac & Feng, 2016). The neuropsychological assessment of these cognitive abilities benefits from an ecologically valid approach to better understand the quality of an individual's everyday functioning (Chaytor & Schmitter-Edgecombe, 2003). Ecological validity increases the probability that an individual's cognitive performance will replicate how they will respond in real-life situations (Bailey, Henry, Rendell, Phillips, & Kliegel, 2010; Burgess *et al.*, 2006; Chaytor & Schmitter-Edgecombe, 2003). Indeed, ecological tests have been found to be better predictors of real-world memory and attention (Higginson, Arnett, & Voss, 2000), executive functioning (e.g., multi-tasking, planning and mental flexibility; Burgess, Alderman, Evans, Emslie, & Wilson, 1998) and prospective memory abilities (e.g., remembering to initiate a planned action in the future; Haines *et al.*, 2019; Phillips, Henry, & Martin, 2012). Several laboratory-based test batteries that simulate real life tasks exist in the neuropsychological literature including those assessing attention (e.g., Test of Everyday Attention, TEA; Robertson, Ward, Ridgeway, & Nimmo-Smith, 1996), memory (e.g., Rivermead Behavioral Memory Test–III, RBMT–III; Wilson, Cockburn, & Baddeley, 2008), executive abilities (e.g., Behavioral Assessment of Dysexecutive Syndrome, BADS; Wilson, Alderman, Burgess, Emslie, & Evans, 1997) and prospective memory (e.g., Cambridge Prospective Memory Test, CAMPROMPT; Wilson, 2005).

Yet, such neuropsychological test batteries tend to incorporate simple, static stimuli within a highly controlled environment and do not fully resemble the complexity of real-life situations (Parsons, 2015; Rand, Rukan, Weiss, & Katz, 2009). Attempts to provide better assessments of everyday abilities have involved assessments in real-life settings such as performing errands in a shopping center or a pedestrianized street (e.g., Garden, Phillips, &

MacPherson, 2001; Shallice & Burgess, 1991). However, these cannot be standardized for use in other clinics or laboratories, while they are not feasible for some populations (e.g., psychiatric patients, stroke patients with paresis or paralysis), they are time-consuming and expensive, they require participant transport and consent from local businesses and they lack experimental control over the external situation (e.g., Elkind, Rubin, Rosenthal, Skoff, & Prather, 2001; Logie, Trawley, & Law, 2011; Parsons, 2015; Rand *et al.*, 2009).

The use of technology such as video recordings of real-world locations and non-immersive virtual environments (Farrimond, Knight, & Titov, 2006; McGeorge *et al.*, 2001; Paraskevaides *et al.*, 2010) have also been considered to simulate real-life situations. Non-immersive virtual reality (VR) tests such as the Edinburgh Virtual Errands Test (EVET; Logie *et al.*, 2011), the Jansari Assessment of Executive Function (Jansari *et al.*, 2014), the Virtual Multiple Errands Test (VMET) within the Virtual Mall (VMall; Rand *et al.*, 2009) and the Virtual Reality Shopping Task (Canty *et al.*, 2014) attempt to simulate real-life tasks and are considered more cost-effective, require less administration time, have greater experimental control and can be easily be adapted for other clinical or research settings (Parsons, McMahan, & Kane, 2018; Werner & Korczyn, 2012; Zygouris & Tsolaki, 2015). Non-immersive VR tests can also offer automated scoring and standardized administration, enabling clinicians and researchers to administer these tests with only limited training. Finally, some non-immersive VR tests also offer shorter versions of the test that focus on the assessment of specific cognitive functions (Parsons *et al.*, 2018; Werner & Korczyn, 2012; Zygouris & Tsolaki, 2015).

However, the user interface and procedure of non-immersive VR tests can be challenging for individuals without gaming backgrounds (Parsons *et al.*, 2018; Zaidi, Duthie, Carr, & Maksoud, 2018), especially for older adults and clinical populations such as individuals with mild cognitive impairment or Alzheimer’s disease (Werner & Korczyn, 2012; Zygouris & Tsolaki, 2015). Immersive VR tests, which share the same advantages as non-immersive ones, may overcome these challenges (Rizzo, Schultheis, Kerns, & Mateer, 2004; Bohil, Alicea, & Biocca, 2011; Parsons, 2015; Teo *et al.*, 2016). In addition, individuals without gaming experience have been found to perform better in immersive VR environments due to the first-person perspective and ergonomic/naturalistic interactions that are proximal to real-life actions (Zaidi *et al.*, 2018). Also, while VR tests have in the past resulted in VR-induced symptoms and effects (VRISE) such as nausea, dizziness, disorientation, fatigue, or instability (Bohil *et al.*, 2011; de Franca & Soares, 2017; Palmisano, Mursic, & Kim, 2017), which compromise neuropsychological (Mittelstaedt, Wacker, & Stelling, 2018; Nalivaiko, Davis, Blackmore, Vakulin, & Nesbitt, 2015; Nesbitt, Davis, Blackmore, & Nalivaiko, 2017) and neuroimaging data (Arafat, Ferdous, & Quarles, 2018; Gavgani *et al.*, 2018; Toschi *et al.*, 2017), certain contemporary VR head-mounted displays (HMDs) and VR software with naturalistic and ergonomic interactions and navigation within the virtual environment reduce or show no symptoms of VRISE (see Kourtesis, Collina, Doumas, & MacPherson, 2019a). Lastly, immersive VR has been found to provide deeper immersion in the virtual environment than non-immersive VR; deeper immersion has been found to induce substantially less adverse VRISE (Kourtesis, Collina, Doumas, & MacPherson, 2019b; Weech, Kenny, & Barnett-Cowan, 2019).

We recently developed the Virtual Reality Everyday Assessment Lab (VR-EAL) to create an immersive virtual environment that simulates everyday tasks proximal to real-life to assess prospective memory, episodic memory (immediate and delayed recognition), executive functions (i.e., multitasking and planning) and selective visual, visuospatial and auditory attention (Kourtesis, Korre, Collina, Doumas, & MacPherson, 2020). In the VR-EAL, individuals are exposed to alternating tutorials (practice trials) and storyline tasks (assessments) to allow them to become familiarized with both the immersive VR technology and the specific controls and procedures of each VR-EAL task. Moreover, VR-EAL offers also a shorter version (i.e., scenario) where only episodic memory, executive function, selective visual attention, and selective visuospatial attention are assessed. Also, the examiner can opt to simply assess a specific cognitive function, where the examinee will go through the generic tutorial, the specific tutorial for this task, and the storyline task that assess the chosen cognitive function (e.g., selective visual attention).

VR-EAL endeavors to be the first immersive VR neuropsychological battery of everyday cognitive functions. Our previous work has shown that the VR-EAL does not induce VRISE (Kourtesis *et al.*, 2020). However, we have yet to demonstrate the validity of the VR-EAL as a neuropsychological tool. In the current study, the full version of the VR-EAL was administered to participants and compared with existing paper-and-pencil neuropsychological tests to assess the construct validity of the VR-EAL. We also aimed to replicate our previous findings that the VR-EAL does not induce VRISE, using the virtual reality neuroscience questionnaire (VRNQ; Kourtesis *et al.*, 2019b). Finally, comparisons between the VR-EAL and neuropsychological paper-and-pencil tests were conducted in terms of verisimilitude (i.e., ecological validity), pleasantness, and administration time.

Methods

Participants

Participants were recruited via social media and the internal mailing list of the University of Edinburgh. Forty-one participants (21 females) aged between 18 and 45 years (M = 29.15, SD = 5.80) were recruited: 18 considered themselves to be gamers (7 females) and 23 (14 females) considered themselves to be non-gamers. The mean education of the group was 13.80 years (SD = 2.36, range = 10-16). The study was approved by the

Philosophy, Psychology and Language Sciences Research Ethics Committee of the University of Edinburgh. Written informed consent was obtained from each participant. All participants received verbal and written instructions regarding the procedures, possible adverse effects of immersive VR (e.g., VRISE), utilization of the data, and general aims of the study.

Materials

Hardware. An HTC Vive HMD with two lighthouse stations for motion tracking and two HTC Vive wands with six degrees of freedom (6DoF) for navigation and interactions within the virtual environment were implemented in accordance with our previously published technological recommendations for immersive VR research (Kourtesis *et al.*, 2019a). The spatialized (bi-aural) audio was facilitated by a pair of Senhai Kotion Each G9000 headphones. The size of the VR area was $5m^2$, which facilitates an adequate space for immersion and naturalistic interaction within virtual environments (Borrego, Latorre, Alcañiz, & Llorens, 2018). The HMD was connected to a laptop with an Intel Core i7 7700HQ 2.80GHz processor, 16 GB RAM, a 4095MB NVIDIA GeForce GTX 1070 graphics card, a 931 GB TOSHIBA MQ01ABD100 (SATA) hard disk, and Realtek High Definition Audio.

VR-EAL. VR-EAL attempts to assess everyday cognitive functioning by assessing prospective memory, episodic memory (i.e., immediate and delayed recognition), executive functioning (i.e., planning, multitasking) and selective visual, visuospatial and auditory (bi-aural) attention within a realistic immersive VR scenario lasting around 60 minutes (Kourtesis *et al.*, 2020). See Table 1 and Figures 1 and 2 for a summary of the VR-EAL tasks assessing each cognitive ability. See Table 2 for the administration procedures and scoring of the VR-EAL tasks.

- Insert Table 1 around here -

Prospective memory. Comparable to the CAMPROMPT, VR-EAL considers both event-based and time-based prospective memory tasks. In the event-based tasks, the participant should remember to perform a prospective memory action when a particular event occurs (e.g., take medicines after breakfast). In the time-based tasks, the examinee should remember to perform a planned action at a specific time (e.g., call Rose at 12 pm).

Episodic memory. Both immediate and delayed episodic memory are assessed. Firstly, the participant needs to memorize a shopping list which is presented audio-visually. Immediately after the presentation of the list, the participant is presented with 30 items and should visually recognize and select the 10 items from the shopping list (immediate recognition). Participants are then expected to choose the items from the list when they arrive at the supermarket approximately 20 minutes later (delayed recognition).

Executive Function: Planning. Planning ability is assessed by asking participants to draw his or her route around the city (e.g., visiting the bakery, supermarket, library, and returning home) on a 3D interactive board.

Executive Function: Multitasking. Multitasking is examined using a cooking task, where the participant should prepare and serve his or her breakfast (e.g., sausages, omelet, and a cup of tea/coffee) and place a chocolate pie in the oven.

Selective visuospatial attention. Visuospatial attention is assessed by asking the participant to find and collect 6 specific items (i.e., a mobile phone, a £50 note, a library card, the flat keys, a red book, and car keys) in the living-room. A reminder of these items remains on the wall (i.e., the items are displayed as 3D objects with labels). However, there are also distractors (i.e., magazines, books, a remote control, a notebook, a pencil, a chessboard, and a bottle of wine) in the room.

Selective visual attention. Visual attention is measured while the participant is seated as a passenger in a car next to a driver. The participant should identify all the targets (i.e., 16 posters of a radio station) on both sides of the road, while s/he needs to avoid any distractors (i.e., 8 posters that are a different shape and 8 posters with a different background color).

Selective auditory attention. Auditory attention is also examined while the participant is seated as a passenger next to a driver. The participant should detect all the target sounds (i.e., 16 bell sounds) presented on both sides of the road, while avoiding the distractor sounds (i.e., 8 high pitch bells, and 8 dongs).

For a full description of the VR-EAL's scenarios, tasks, and scoring, see Kourtesis *et al.* (2020). Also, a brief video recording of the VR-EAL may be accessed at this hyperlink: https://www.youtube.com/watch?v=IHEIvS37Xy8&t= .

- Insert Figures 1 and 2 and Table 2 around here -

Paper-and-Pencil Tests. Established ecologically valid paper-and-pencil test batteries (i.e., CAMPROMPT; RBMT-III; BADS; TEA) were selected to match the equivalent VR-EAL tasks and examine their ecological and construct validity. However, two less ecologically valid neuropsychological tests were also included to assess the validity of the VR-EAL's visuospatial attention and multitasking tasks, since there were not any published ecologically valid paper-and-pencil tasks for the assessment of these cognitive functions.

Prospective Memory. The CAMPROMPT was administered to evaluate prospective memory using six prospective memory tasks (Wilson, 2005). Three tasks are event-based, and three are time-based. The participant is required to perform several distractor tasks (e.g., word-finder puzzles and general knowledge quizzes and questions) for 20 minutes, as well as remember to perform the prospective memory tasks (e.g., when the participant faces a question which includes the word “EastEnders”, s/he needs to give a book to the examiner). The utilization of reminding strategies (e.g., taking notes) is permitted to aid the participant to remember when and how to perform the prospective memory tasks. The CAMPROMPT provides three scores: a total score (out of 36), an event-based score (out of 18), and a time-based score (out of 18).

Episodic Memory. Two subtests from the RBMT-III (Wilson *et al.*, 2008) were administered to assess episodic memory. The recall tasks were opted since they offer two scores (immediate recall, delayed recall), while the recognition tasks provide a score only for delayed recognition. The immediate and delayed story recall tasks were used to match the VR-EAL’s immediate and delayed recognition tasks. The participant listens to a story from a newspaper read aloud by the examiner. The participant should recall the story immediately (immediate recall out of 21) and after approximately 20 minutes (delayed recall out of 21).

Executive Function: Planning. The Key Search task from the BADS (Wilson *et al.*, 1997) was utilized as a test of planning (Wilson, Evans, Emslie, Alderman, & Burgess, 1998). While the Key Search task assesses planning ability, it also involves other aspects of executive function (e.g., problem-solving, and monitoring of behavior; Wilson *et al.*, 1998). The participant should draw his or her route to find lost keys in a field. The quality of the route (e.g., whether it covers the whole field) and the time taken to draw it are considered in the scoring (max score = 16).

Executive functioning. The Color Trails Test (CTT; D’Elia, Satz, Uchiyama, & White, 1996) was administered to assess processing speed and executive functioning. CTT is a non-alphabetical adaptation (i.e., colors and numbers) of the Trail Making Test (Reitan, &

Wolfson, 1993). CTT has two tasks (i.e., CTT-1 and CTT-2), where the participant must draw a line to connect consecutive numbers. In CTT-1, the numbers in the sequence are in a single color. Comparable to the TMT-A, CTT-1 assesses processing speed. In CCT-2, the numbers are displayed in two colors and the examinee alternates between the two colors for each number in the sequence. Comparable to the TMT-B, CTT-2 assesses task-switching, as well as inhibition and visual attention (D'Elia *et al.*, 1996). The CTT was chosen to assess the validity of the VR-EAL's multitasking task, and these aspects of executive functioning have been found central in everyday multitasking (Logie *et al.*, 2011). Furthermore, the time to complete in seconds is taken as the score for CTT-1 and CTT-2, and the difference between the two scores (i.e., CTT-2 minus CTT-1) is considered an index of executive function.

Selective visual attention. The Ruff 2 and 7 Selective Attention Test (RSAT; Ruff, Niemann, Allen, Farrow, & Wylie, 1992) was used to assess selective visual attention. The participant is asked to identify target numbers (i.e., 2s and 7s) and ignore the distractors (either numbers or letters) in the block. The examinee is required to implement two different strategies for each type of block; an automatic selection of 2s and 7s for the blocks with letter-distractors, and a controlled detection of 2s and 7s for the blocks with number-distractors. The RSAT produces two scores: a detection speed score (out of 80) and a detection accuracy score (out of 59). The scores consider the number of detected 2s and 7s, as well as, the number of misses and errors. The RSAT was opted to match the VR-EAL selective visuospatial attention task because it requires different scanning strategies, shifting of attention to another block, and considers the number of misses and mistakes.

Selective visual attention. The Map task from the TEA (Robertson *et al.*, 1994) was administered to assess selective visual attention (i.e., the ability to detect visual targets, while disregarding similar visual distractors). The participant should find as many as possible restaurant symbols (version A) or gas station symbols (version B) on a map of Philadelphia

(USA) within two minutes. The total score out of 80 corresponds to the number of symbols detected overall, while one subscore corresponds to the number of symbols found in the first minute, and the other subscore refers to the number of symbols detected in the second minute.

Selective auditory attention. The Elevator Counting with Distraction task of the TEA (Robertson *et al.*, 1994) was administered, which measures auditory selective attention (i.e., the ability to select target sounds, while ignoring competitive auditory distractors). In each trial, the participant listens to different sounds (beeps), where s/he needs to count the number of normal pitched beeps (i.e., targets) and disregard the high pitched and low-pitched beeps (i.e., distractors). The total score is the number of correct responses across the 10 trials (max score = 10).

Questionnaires. The VRNQ was administered to assess the quality of the VR-EAL and the intensity of VRISE. A survey questionnaire was administered to evaluate the gaming and VR experience of the participants (see Supplementary Material – Figure 1). A comparison questionnaire (two versions; i.e., VR and paper-and-pencil) was administered to examine the participants' views on the pleasantness and ecological validity of the tests performed (see Supplementary Material – Figures 2 and 3).

VR software quality and VRISE. The VRNQ is a 1–7 Likert scale questionnaire comprising 20 questions in total; 5 questions are pertinent to each of the 4 domains (i.e., user experience, game mechanics, in-game assistance and VRISE) (Kourtesis *et al.*, 2019b). The assessed VRISE are nausea, dizziness, disorientation, fatigue, and instability. VRNQ produces a total score out of 140 and a subscore out of 35 for each domain. The parsimonious cut-offs of VRNQ were used to assess the suitability of VR-EAL (Kourtesis *et al.*, 2020).

Gaming and VR experience. The questionnaire (Likert scale 1-7) contains two questions regarding the weekly frequency of game playing and VR technology use, and two questions pertinent to the ability to play games and VR technologies use.

Verisimilitude and pleasantness. There were two separate versions of the comparison questionnaire with a Likert scale ranging from 1 to 7. There was one version for the VR-EAL tasks (see Supplementary Material – Figure 2), and another for the paper-and-pencil tests (see Supplementary Material – Figure 3). Both versions had the same two questions referring to the level of enjoyment (e.g., 1-highly unpleasant, 7-highly pleasant) and verisimilitude (e.g., 1- totally different from the tasks in daily life, 7-nearly identical to the tasks in daily life) of the tasks. For each version of the questionnaire, the maximum score was 14.

Procedure

Participants individually attended both the VR session and the paper-and-pencil session; the order was pseudorandomized across participants. In the VR session, participants participated in an induction session to introduce them to the HMD and controllers (i.e., HTC Vive and 6DoF wands-controllers) prior to immersion. After completion of VR-EAL, participants completed the VRNQ and the VR versions of the comparison questionnaire (i.e., to assess pleasantness and verisimilitude). During the paper-and-pencil session, participants completed the paper-and-pencil comparison questionnaires (i.e., pleasantness and verisimilitude) after each test. The duration of each session was timed using a stopwatch.

Statistical Analyses

The Bayesian factor (BF_{10}) was used for assessing statistical inference. The BF_{10} threshold ≥ 10 was set for statistical inference in all analyses, which indicates strong evidence in favor of the H1 (Marsman & Wagenmakers, 2017; Rouder & Morey, 2012; Wetzels &

Wagenmakers, 2012) and corresponds to a p-value < 0.01 (e.g., BF_{10} = 10) (Bland, 2015; Cox & Donnelly, 2011; Held & Ott, 2018). BF_{10} is considered substantially more parsimonious than the p-value in evaluating the evidence against the H0 (Bland, 2015; Cox & Donnelly, 2011; Held & Ott, 2018), especially when evaluating the evidence of H1 against H0 in small sample sizes (Held & Ott, 2018), as in the present study. Notably, BF_{10} allows evidence in either direction (i.e., towards H1 and H0), and its measurement of evidence is insensitive to the stopping rule, which substantially mitigates the issue of multiple comparisons and generates reliable and more generalizable results (Dienes, 2016; Marsman & Wagenmakers, 2017; Wagenmakers *et al.*, 2018b).

Bayesian Pearson correlational analyses were conducted to examine associations between age, years of education, VR experience, gaming experience, and performance on the VR-EAL and paper-and-pencil tasks. Similarly, Bayesian Pearson correlational analyses were performed to assess construct validity for the entire VR-EAL and convergent validity between the VR-EAL tasks and the paper-and-pencil tasks. Furthermore, Bayesian paired samples t-tests were performed to investigate the differences between VR-EAL and paper-and-pencil tests in terms of verisimilitude, pleasantness, and administration time. Finally, a post hoc analyses for the achieved statistical power of the Bayesian Pearson's correlations and Bayesian paired samples t-tests were performed using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009). All Bayesian analyses were performed using JASP (Version 0.8.1.2) (JASP Team, 2018).

Results

The descriptive statistics of the sample performing the VR-EAL, the paper-and-pencil tests and questionnaires are displayed in Table 3.

- Insert Table 3 around here -

Correlations between demographics and performance

No significant correlations were found between age, education, VR experience, gaming experience, or performance on any of the paper-and-pencil tests or the VR-EAL tasks. The only significant correlations were observed between gaming experience and VR experience, VR experience and the VR session duration, gaming experience and the VR session duration, gaming experience and the duration of the paper-and-pencil testing session, and the duration of the VR session and the paper-and-pencil session (see Table 4).

- Insert Table 4 around here -

Convergent and construct validity of the VR-EAL

The VR-EAL scores were significantly positively correlated with their equivalent scores on the paper-and-pencil tests (see Table 5). These results support the convergent validity of the VR-EAL tasks, as well as the construct validity of the VR-EAL as an immersive VR neuropsychological battery.

- Insert Table 5 around here -

Quality of VR-EAL and VRISE: VRNQ

The median of the VRNQ total score for VR-EAL was 128, which is substantially above the parsimonious cut-off of 120 (maximum score = 140). The medians of the VRNQ domains (i.e., user experience, game mechanics, in-game assistance and VRISE) were between 31 and 33, again above their respective parsimonious cut-offs of 30 (maximum score = 35). Notably, the medians for all the individual VRISE items (i.e., nausea, dizziness, disorientation, fatigue, and instability) were 7 (i.e., absent feeling), except for fatigue, which was 6 (i.e., very mild feeling). No participant reported a VRISE subscore less than 5 (i.e., mild feeling).

Comparison of the testing experience between VR-EAL and paper-and-pencil tests

The median for enjoyment level was 6 (very pleasant) for the VR-EAL and 5 (pleasant) for the paper-and-pencil assessments (see Figure 3). The median for verisimilitude was 6 (i.e., very similar to everyday life) for VR-EAL, 4 (neither similar nor dissimilar to everyday life) for the ecologically validity tests, and 3 (dissimilar to everyday life) for the remaining paper-and-pencil tests (see Figure 3). The Bayesian t-tests demonstrated significant differences between the VR-EAL and paper-and-pencil tests, where the VR-EAL is rated significantly more pleasant and ecologically valid (i.e., verisimilitude) than the paper-and-pencil tests (see Table 6). In addition, the VR session was substantially shorter than the paper-and-pencil session (see Table 6).

- Insert Figure 3 and Table 6 around here -

Discussion

The VR-EAL was devised to assess cognitive functions (i.e., prospective memory, episodic memory, executive functions, and attentional processes) that are central to everyday functioning. Being an immersive VR research/clinical software, the VR-EAL aims to increase the likelihood that individuals' performance will replicate how they will act in real-life situations (Higginson *et al.*, 2000; Chaytor & Schmitter-Edgecombe, 2003; Phillips *et al.*, 2012; Rosenberg, 2015; Mlinac & Feng, 2016; Haines *et al.*, 2019). In the current study, we attempted to provide construct and convergent validity for the VR-EAL tasks. Indeed, we demonstrated that all VR-EAL tasks significantly correlated with their corresponding paper-and-pencil tasks. Notably, the VR-EAL tasks correlated with their corresponding ecologically valid neuropsychological test (i.e., TEA, RBMT, BADS, and CAMPROMPT). Therefore, the VR-EAL appears to be an effective and ecologically valid tool for the assessment of everyday cognitive functioning, which can be used for clinical and research purposes. Importantly, the VR-EAL is a highly immersive and ergonomic VR neuropsychological battery; immersive VR provides a more ecological valid experience than non-immersive VR (Weech, Kenny, & Barnett-Cowan, 2019) and ergonomic interactions benefit non-gamers as their performance is comparable to gamers (Zaidi *et al.*, 2018).

Previous studies examining the ecological validity of other VR neuropsychological tools have not considered users' perceptions of the task's ecological validity (e.g., Canty *et al.*, 2014; Jansari *et al.*, 2014; Logie *et al.*, 2011; Rand *et al.*, 2009;). Therefore, a further advantage of VR-EAL is that the participants rated it as more similar to the tasks that they perform in their daily life (i.e., ecologically valid) than all tests in the paper-and-pencil neuropsychological battery and the group of well-established ecological valid tests (i.e., CAMPROMPT test, RBMT-Story Recall, BADS-Key Search, TEA-Map, and TEA-Elevator Counting with Distraction). Furthermore, the VR-EAL tasks were individually compared to their corresponding paper-pencil test, where the results postulated that the VR-EAL tasks are significantly more ecologically valid than the equivalent paper-pencil tests. Also, as far as we are aware, our study is the first to compare the pleasantness of the testing experience between immersive VR and paper-and-pencil tests. Here, the full-version of the VR-EAL was also considered by the participants to be a more pleasant testing experience than the paper-and-pencil neuropsychological battery. Furthermore, the duration of the entire VR session (i.e., the induction and performance of VR-EAL) was considerably shorter than the administration time for the paper-and-pencil neuropsychological battery. Therefore, the VR-EAL emerges as substantially more enjoyable and ecologically valid testing experience with a significantly shorter administration time in comparison with the equivalent paper-and-pencil neuropsychological battery.

Age and education did not correlate with performance on the VR-EAL or the paper-and-pencil tests. While the paper-and-pencil scores were adjusted for age and education, the VR-EAL scores were not. Therefore, the VR-EAL may have the advantage that performance is not dependent on age or education. However, this needs to be further investigated in a larger and more diverse population, as the population of this study predominantly comprised younger adults aged 18 to 45 years with a relatively high level of education (i.e., 10-16 years).

Gaming experience strongly and positively associated with VR experience, indicating that gamers are also more experienced immersive VR users. Also, VR and gaming experience were both negatively correlated with the duration of the VR session, where more experienced gamers complete the assessment faster than non-gamers. Interestingly, however, gaming experience was also correlated with the duration of the paper-and-pencil session, indicating that gamers complete the paper-and-pencil assessment faster than non-gamers. Finally, the duration of the VR session was correlated significantly with that of the paper-and-pencil session, which also indicates that the speed of performing tasks affects the duration of both types of tasks (i.e., immersive VR and paper-and-pencil). Our findings are aligned with the relevant literature where gamers have been found to have enhanced perceptual processing speed (Anguera *et al.*, 2013; Dye, Green, & Bavelier, 2009; Kowal, Toth, Exton, &

Campbell, 2018). However, in our sample, the gaming ability was not associated with the performance on the cognitive tests, indicating that gaming ability is not linked with an improved overall cognition, which is also in line with the relevant literature (Kowal *et al.*, 2018).

Another aim of the study was to provide immersive VR software for clinical and research use that has minimal VRISE, since adverse symptomology associated with VR can significantly decrease participants' reaction times and overall cognitive performance (Nalivaiko *et al.*, 2015; Nesbitt *et al.*, 2017; Mittelstaedt *et al.*, 2018). Albeit that the incidence of VRISE is more frequent in immersive VR, these symptoms are also highly frequent in non-immersive VR (Sharples, Cobb, Moody, & Wilson, 2008). However, the examination and report of VRISE has not been considered in non-immersive VR studies of neuropsychological tools for clinical and research purposes (e.g., Canty *et al.*, 2014; Jansari *et al.*, 2014; Logie *et al.*, 2011; Rand *et al.*, 2009). Similarly, the examination of VRISE is under-reported or not examined in immersive VR studies of neuropsychological tools (Kourtesis *et al.*, 2019a).

In contrast, the examination and report of VRISE was central in our endeavour to scrutinise the suitability of VR-EAL as a neuropsychological tool for research and clinical purposes. Our current findings replicate those of our previous work where VR-EAL did not induce VRISE in participants (Kourtesis *et al.*, 2020). In this study, VR-EAL exceeded the parsimonious cut-offs for the VRNQ scores (total score, user experience, game mechanics, in-game assistance, and VRISE). The outcomes of VRNQ hence postulate that VR-EAL is a suitable VR software for implementation in research and clinical settings, without inducing VRISE. On all VRISE items, except fatigue, there was an absence of adverse symptoms. Participants reported only very mild feelings of fatigue albeit that this was an expected outcome since the duration of VR-EAL was around 60 minutes. However, fatigue was equally present during the paper-and-pencil session (80 minutes).

This study also has some limitations. The sample was moderately small (N = 41), though, every statistical analysis displayed a substantially robust statistical power (>90%). Moreover, as the current study is the first to provide validity for the VR-EAL, it was only administered to younger but not older adults. Yet, the eventual aim is to use the VR-EAL to assess cognitive impairments in healthy aging and dementias (Anderson & Craik, 2017) or attention-deficit/hyperactivity disorder and autism (Karalunas *et al.*, 2018). Future work should examine the performance and experiences of different clinical populations performing the VR-EAL to provide further evidence for the clinical utility of VR-EAL for assessing everyday cognitive functioning.

In summary, this study provides evidence supporting the validation of VR-EAL as an effective neuropsychological tool with enhanced ecological validity for the assessment of everyday cognitive functioning. In addition, the VR-EAL does not seem to induce VRISE (i.e., cybersickness). Therefore, our preliminary findings support the VR-EAL as an immersive VR assessment tool that has the potential to be implemented in both research and clinical settings in the future.

Acknowledgments

The authors declare no conflicts of interest. The current study did not receive any financial support or grants.

References

Anderson, N. D., & Craik, F. I. (2017). 50 years of cognitive aging theory. *The Journals of Gerontology: Series B*, 72(1), 1-6.

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... & Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, 501(7465), 97.

Arafat, I. M., Ferdous, S. M. S., & Quarles, J. (2018, March). Cybersickness-Provoking Virtual Reality Alters Brain Signals of Persons with Multiple Sclerosis. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 1-120). IEEE.

Bailey, P. E., Henry, J. D., Rendell, P. G., Phillips, L. H., & Kliegel, M. (2010). Dismantling the "age–prospective memory paradox": The classic laboratory paradigm simulated in a naturalistic setting. *Quarterly Journal of Experimental Psychology*, 63, 646–652.

Bland, M. (2015). *An introduction to medical statistics.* Oxford, UK: Oxford University Press.

Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience*, 12(12), 752-762.

Borrego, A., Latorre, J., Alcañiz, M., & Llorens, R. (2018). Comparison of Oculus Rift and HTC Vive: feasibility for virtual reality-based exploration, navigation, exergaming, and rehabilitation. *Games for Health Journal*, 7(3), 151-156.

Burgess, P. W., Alderman, N., Evans, J., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, 4(6), 547-558.

Burgess, P. W., Alderman, F., Frobes, C., Costello, A., Coates, L. M., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, 12, 194–209.

Canty, A. L., Fleming, J., Patterson, F., Green, H. J., Man, D., & Shum, D. H. (2014). Evaluation of a virtual reality prospective memory task for use with individuals with severe traumatic brain injury. *Neuropsychological Rehabilitation*, 24(2), 238-265.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, 13(4), 181-197.

Cox, D. R., & Donnelly, C. A. (2011). Principles of applied statistics. *Cambridge University Press*.

de França, A. C. P., & Soares, M. M. (2017, July). Review of Virtual Reality Technology: An Ergonomic Approach and Current Challenges. *In International Conference on Applied Human Factors and Ergonomics* (pp. 52-61). Springer, Cham.

D'Elia, L. F., Satz, P., Uchiyama, C. L., & White, T. (1996). *Color Trails Test: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78-89.

Dye, M. W., Green, C. S., & Bavelier, D. (2009). Increasing speed of processing with action video games. *Current Directions in Psychological Science*, 18(6), 321-326.

Elkind, J. S., Rubin, E., Rosenthal, S., Skoff, B., & Prather, P. (2001). A simulated reality scenario compared with the computerized Wisconsin card sorting test: An analysis of preliminary results. *CyberPsychology & Behavior*, 4(4), 489–496.

Farrimond, S., Knight, R. G., & Titov, N. (2006). The effects of aging on remembering intentions: Performance on a simulated shopping task. *Applied Cognitive Psychology*, 20, 533–555.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149-1160.

Garden, S., Phillips, L. H. & MacPherson, S. E. (2001). Mid-life aging, open-ended planning, and laboratory measures of executive function. *Neuropsychology*, 15(4), 472-482.

Gavgani, A. M., Wong, R. H., Howe, P. R., Hodgson, D. M., Walker, F. R., & Nalivaiko, E. (2018). Cybersickness-related changes in brain hemodynamics: A pilot study comparing transcranial Doppler and near-infrared spectroscopy assessments during a virtual ride on a roller coaster. *Physiology & Behavior*, 191, 56-64.

Haines, S., Shelton, J., Henry, J., Terrett, G., Vorwerk, T., & Rendell, P. (2019, February 25). Prospective Memory and Cognitive Aging. *Oxford Research Encyclopedia of Psychology*. Retrieved 7 Dec. 2019, from https://oxfordre.com/psychology/view/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-381.

Held, L., & Ott, M. (2018). On p-values and Bayes factors. *Annual Review of Statistics and Its Application*, 5, 393-419.

Higginson, C. I., Arnett, P. A., & Voss, W. D. (2000). The ecological validity of clinical tests of memory and attention in multiple sclerosis. *Archives of Clinical Neuropsychology*, 15(3), 185-204.

Jansari, A. S., Devlin, A., Agnew, R., Akesson, K., Murphy, L., & Leadbetter, T. (2014). Ecological assessment of executive functions: a new virtual reality paradigm. *Brain Impairment*, 15(2), 71-87.

JASP Team (2017). JASP (Version 0.8.1.2) [Computer software].

Karalunas, S. L., Hawkey, E., Gustafsson, H., Miller, M., Langhorst, M., Cordova, M., ... & Nigg, J. T. (2018). Overlapping and distinct cognitive impairments in attention-deficit/hyperactivity and autism spectrum disorder without intellectual disability. *Journal of Abnormal Child Psychology*, 46(8), 1705-1716.

Kourtesis P., Collina S., Doumas L.A.A., & MacPherson S.E. (2019a). Technological competence is a precondition to effectively implement virtual reality head mounted displays in human neuroscience: A technological review and meta-analysis. *Frontiers in Human Neuroscience*, 13, 342.

Kourtesis P., Collina S., Doumas L.A.A., & MacPherson S.E. (2019b). Validation of the Virtual Reality Neuroscience Questionnaire: Maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Frontiers in Human Neuroscience*, 13, 417.

Kourtesis P, Korre D, Collina S, Doumas LAA, & MacPherson SE (2020) Guidelines for the Development of Immersive Virtual Reality Software for Cognitive Neuroscience and Neuropsychology: The Development of Virtual Reality Everyday Assessment Lab (VR-EAL), a Neuropsychological Test Battery in Immersive Virtual Reality. *Frontiers in Computer Science*, 1, 12.

Kowal, M., Toth, A. J., Exton, C., & Campbell, M. J. (2018). Different cognitive abilities displayed by action video gamers and non-gamers. *Computers in Human Behavior*, 88, 255-262.

Logie, R. H., Trawley, S., & Law, A. S. (2011). Multitasking: Multiple, domain specific cognitive functions in a virtual environment. *Memory and Cognition,* 39, 1561–1574.

Marsman, M., & Wagenmakers, E. J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, 14(5), 545-555.

McGeorge, P., Phillips, L. H., Crawford, J. R., Garden, S. E., Della Sala, S. D., Milne, A. B., . . . Callender, J. S. (2001). Using virtual environments in the assessment of executive dysfunction. *Presence: Teleoperators and Virtual Environments*, 10(4), 375–383.

Mittelstaedt, J. M., Wacker, J., & Stelling, D. (2018). VR aftereffect and the relation of cybersickness and cognitive performance. *Virtual Reality*, 1-12.

Mlinac, M. E., & Feng, M. C. (2016). Assessment of activities of daily living, self-care, and independence. *Archives of Clinical Neuropsychology*, 31(6), 506-516.

Nalivaiko, E., Davis, S. L., Blackmore, K. L., Vakulin, A., & Nesbitt, K. V. (2015). Cybersickness provoked by head-mounted display affects cutaneous vascular tone, heart rate and reaction time. *Physiology & Behavior*, 151, 583-590.

Nesbitt, K., Davis, S., Blackmore, K., & Nalivaiko, E. (2017). Correlating reaction time and nausea measures with traditional measures of cybersickness. *Displays*, 48, 1-8

Palmisano, S., Mursic, R., & Kim, J. (2017). Vection and cybersickness generated by head-and-display motion in the Oculus Rift. *Displays*, 46, 1-8.

Paraskevaides, T., Morgan, C. J. A., Leitza, J. R., Bisby, J. A., Rendell, P. G., & Curran, H. V. (2010). Drinking and future thinking. Acute effects of alcohol on prospective memory and future simulation. *Psychopharmacology*, 208, 301–308.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, 9.

Parsons, T. D., McMahan, T., & Kane, R. (2018). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. *The Clinical Neuropsychologist*, 32(1), 16-41.

Phillips, L. H., Henry, J. D., & Martin, M. (2008). Adult aging and prospective memory: The importance of ecological validity. In M. Kliegel, M. A. McDaniel, & G. O. Einstein

(Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (p. 161–185). Taylor & Francis Group/Lawrence Erlbaum Associates.

Rand, D., Rukan, S. B. A., Weiss, P. L., & Katz, N. (2009). Validation of the Virtual MET as an assessment tool for executive functions. *Neuropsychological Rehabilitation*, 19(4), 583-602.

Reitan, R., & Wolfson, D. (1993). The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation. Tucson, AZ: Neuropsychology Press.

Rizzo, A. A., Schultheis, M., Kerns, K. A., & Mateer, C. (2004). Analysis of assets for virtual reality applications in neuropsychology. *Neuropsychological Rehabilitation,* 14(1-2), 207-239.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). The test of everyday attention (TEA). *Bury St. Edmunds, UK: Thames Valley Test Company*, 197-221.

Rosenberg, L. (2015). The associations between executive functions' capacities, performance process skills, and dimensions of participation in activities of daily life among children of elementary school age. *Applied Neuropsychology: Child*, 4(3), 148-156.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6), 877-903.

Ruff, R. M., Niemann, H., Allen, C. C., Farrow, C. E., & Wylie, T. (1992). The Ruff 2 and 7 selective attention test: a neuropsychological application. *Perceptual and Motor Skills*, *75*(3_suppl), 1311-1319.

Shallice, T., & Burgess, P. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain,* 114, 727–741.

Sharples, S., Cobb, S., Moody, A., and Wilson, J. R. (2008). Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (HMD), desktop and projection display systems. *Displays,* 29, 58–69.

Teo, W. P., Muthalib, M., Yamin, S., Hendy, A. M., Bramstedt, K., Kotsopoulos, E., ... & Ayaz, H. (2016). Does a combination of virtual reality, neuromodulation and neuroimaging provide a comprehensive platform for neurorehabilitation?–A narrative review of the literature. *Frontiers in Human Neuroscience*, 10, 284

Toschi, N., Kim, J., Sclocco, R., Duggento, A., Barbieri, R., Kuo, B., & Napadow, V. (2017). Motion sickness increases functional connectivity between visual motion and nausea-associated brain regions. *Autonomic Neuroscience*, 202, 108-113.

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: a review. *Frontiers in Psychology*, 10, 158.

Werner, P., & Korczyn, A. D. (2012). Willingness to use computerized systems for the diagnosis of dementia: testing a theoretical model in an Israeli sample. *Alzheimer Disease & Associated Disorders*, 26(2), 171-178.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic bulletin & review*, *19*(6), 1057-1064.

Wilson, B. A. (2005). *The Cambridge prospective memory test: CAMPROMPT*. London: Harcourt Assessment.

Wilson, B. A., Cockburn, J., & Baddeley, A. (2008). *The Rivermead behavioural memory test*. Bury St Edmunds, UK: Thames Valley Test Company.

Wilson, B. A., Alderman, N., Burgess, P. W., Emslie, H., & Evans, J.J. (1996). Behavioural Assessment of the Dysexecutive Syndrome (BADS). Bury St. Edmunds, UK, Thames Valley Test Company.

Wilson, B. A., Evans, J. J., Emslie, H., Alderman, N., & Burgess, P. (1998). The development of an ecologically valid test for assessing patients with a dysexecutive syndrome. *Neuropsychological Rehabilitation*, 8(3), 213-228.

Zaidi, S. F. M., Duthie, C., Carr, E., & Maksoud, S. H. A. E. (2018, December). Conceptual framework for the usability evaluation of gamified virtual reality environment for non-gamers. In *Proceedings of the 16th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry* (p. 13). ACM.

Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: a review. *American Journal of Alzheimer's Disease & Other Dementias®*, 30(1), 13-28.

Table 1. VR-EAL tasks and score ranges

Scene	Cognitive Function	Task	Score Ranges
3	Prospective memory	Write down the notes for the errands.	0 – 6
3	Immediate recognition	Recognising items on the shopping list.	0 – 20
3	Planning	Drawing the route to be taken.	0 – 19
6	Multitasking	Cooking task (preparing breakfast).	0 – 16
6	Prospective memory – event based	Take medication after breakfast.	0 – 6
8	Selective visuospatial attention	Collect items from the living room.	0 – 20
8	Prospective memory – event based	Take the chocolate pie out of the oven.	0 – 6
10	Prospective memory – time based	Call Rose at 10 am.	0 – 6
12	Selective visual attention	Find posters on both sides of the road.	0 – 16
14	Delayed recognition	Recognising items from the shopping list.	0 – 20
15	Prospective memory – time based	Collect the carrot cake from the bakery at 12 pm.	0 – 6
16	Prospective memory – event based	False prompt before going to the library.	-6 – 0
17	Prospective memory – event based	Return the red book to the library.	0 – 6
19	Selective auditory attention	Detect sounds from both sides of the road.	0 – 32
20	Prospective memory – time based	False prompt before going back home.	-6 – 0
21	Prospective memory – event based	Back home, give the extra pair of keys to Alex.	0 – 6
22	Prospective memory – time based	Take the medication at 1pm.	0 – 6

*The tasks are presented in the same order as they are performed within the scenario.

Table 2. VR-EAL task administration and scoring

Task	Scoring
Episodic memory	The user should choose the ten target items (i.e., create the shopping list) from an extensive array of items, which also contains five qualitative distractors (e.g., semi-skimmed milk versus skimmed milk), five quantitative distractors (e.g., 1 kg potatoes versus 2 kg potatoes), and ten false items (e.g., bread, bananas etc.). The user gains 2 points for each correctly chosen item, 1 point for choosing a qualitative or quantitative distractor, and 0 points for the false items. Scores range from 0 to 20.
Planning	The road system comprises 23 street units. When the user selects a unit, 1 point is awarded. The ideal route to visit all three destinations is 15 units; hence, any extra or missing units are subtracted from the total possible score of 15. Up to 4 more points are awarded for the time taken to complete the task. Scores range from 0 to 19.
Multitasking	Scoring relies on the animations from each game object (i.e., the omelet and the sausages). At the beginning of the animation, both items have a reddish (raw) color which gradually turns to either a yellowish (omelet) or brownish (sausages) color, and finally both turn to black (burnt). The score for each pan hence depends on the time that the user removes the pans from the stove and places them on the kitchen worktop. Equally, the score for boiling the kettle is measured in relation to the stage of the audio playback (e.g., the kettle whistles when the water is ready) that the kettle is placed on the kitchen worktop. Scores range from 0 to 16.
Prospective Memory	**Example:** At the end of a scene, the user should press a button to confirm that all the tasks in the scene are completed. If the user has already taken his/her medication (i.e., prospective memory task) before pressing the final button, then the scene ends, and the user receives 6 points. Otherwise, the first prompt appears (i.e., “You Have to Do Something Else”). If the user then follows the prompt and takes their medication, they receive 4 points. If the user presses the final button again, then the second prompt appears (i.e., “You Have to Do Something After Having your Breakfast”). If the user follows this prompt and takes their medication, they receive 2 points. If the user presses the final button again, then the third prompt appears (i.e., “You Have to Take Your Meds”). If the user follows this prompt and takes their medication, they then receive 1 point. If the user represses the final button without ever taking their medication, they get zero points, and the scene ends. Scores range from 0 to 6.
Visuospatial Attention	The user receives 2 point for each target item collected (6 target items). Also, up to 4 points are awarded for the speed of detecting the items. If the user attempts to collect one of the distractors, it counts as an error. Up to 4 points are awarded for the accuracy of detecting items. Scores range from 0 to 20.
Visual Attention	The user is awarded 1 point when a target poster is “spotted” and subtracted 1 point when a distractor poster is “spotted”. Scores range from 0 to 16.
Auditory Attention	**Example:** if the user presses the trigger on the right controller to detect a target sound originating on the right side (i.e., controller and sound on the same side), then s/he gets 2 points. If the user presses the trigger on the right controller to detect a target sound originating on the left side (i.e., controller on the opposite side), s/he gains only 1 point. If the user responds to a distractor sound, irrelevant of its origin or the controller used to respond, 1 point is deducted. Scores range from 0 to 32.

Note: For all measures, higher scores indicate better performance.

Table 3. Descriptive statistics for the VR-EAL, paper-and-pencil tests and questionnaires

	N	Mean (SD)	Range
Gaming Experience	41	6.12 (3.95)	2-13
VR Experience	41	3.29 (1.29)	2-6
Total Time VR-EAL (in minutes)	41	63.95 (7.88)	50-81
Total Time VR Session (in minutes)	41	73.95 (7.88)	60-91
Total Time Paper-Pencil Assessment (in minutes)	41	85.41 (3.97)	76-92
CAMPROMPT – Total Score (max = 36)	41	30.83 (3.49)	24-36
VR-EAL – PM Total Score (max = 48)	41	35.78 (4.73)	24-46
CAMPROMPT - Event Based (max = 18)	41	16.39 (1.63)	12-18
VR-EAL - Total Event Based (max = 24)	41	18.15 (3.26)	8-24
CAMPROMPT - Time Based (max = 18)	41	14.44 (2.66)	10-18
VR-EAL - Time Based (max = 18)	41	11.63 (3.10)	6-18
RBMT - Immediate Recall (max = 21)	41	14.93 (2.24)	10-18
VR-EAL - Immediate Recognition (max = 20)	41	15.51 (1.98)	10-18
RBMT - Delayed Recall (max = 21)	41	15.98 (2.61)	11-21
VR-EAL - Delayed Recognition (max = 20)	41	17.17 (2.42)	12-20
TEA – Map Total Score (max = 80)	41	70.32 (6.87)	52-82
VR-EAL - Selective Visual Attention Accuracy (max = 32)	41	22.98 (3.84)	17-30
RSAT – Accuracy (max = 59)	41	47.51 (7.14)	27-58
VR-EAL - Selective Visual Attention Speed (max = 32)	41	23.61 (3.69)	18-30
RSAT – Speed (max = 80)	41	57.78 (9.39)	33-74
VR-EAL - Selective Visuospatial Attention Total (max = 20)	41	12.00 (2.42)	4-15
VR-EAL - Selective Visuospatial Attention Speed (max = 16)	41	11.90 (1.50)	8-14
VR-EAL - Selective Visuospatial Attention Accuracy (max = 16)	41	12.10 (1.18)	8-13
TEA - Elevator Counting with Distraction (max = 10)	41	9.05 (1.05)	7-10
VR-EAL - Selective Auditory Attention (max = 32)	41	29.56 (3.66)	20-32
BADS – Key Search (max = 16)	41	14.20 (1.47)	10-16
VR-EAL – Planning (max = 19)	41	14.90 (1.51)	11-17
CTT – 1 (max = 80)	41	49.37 (8.65)	32-68
VR-EAL - Cooking Task (max = 16)	41	9.68 (2.57)	2-13
CTT – 2 (max = 80)	41	55.20 (9.94)	27-70

VR-EAL = Virtual Reality Everyday Assessment Lab; CAMPROMPT = Cambridge Prospective Memory Test; RBMT = Rivermead Behavioral Memory Test; TEA = Test of Everyday Attention; BADS = Behavioral Assessment of Dysexecutive Syndrome; CTT = Color Trails Test.

Table 4. Bayesian correlations between users' experience and the sessions' durations.

Correlational Pairs	r	BF_{10}	SP
Gaming experience - VR experience	0.84^{***}	1.72e+10	~ 100%
VR experience - VR session duration	-0.60^{***}	690.55	99%
Gaming experience - VR session duration	-0.55^{***}	136.41	97%
Gaming experience - Paper-and-pencil session duration	-0.45^{***}	12.17	94%
VR session duration - Paper-and-pencil session duration	0.53^{***}	87.22	97%

The alternative hypothesis specifies that the correlation is positive. * $BF_{10} > 10$; ** $BF_{10} > 30$; *** $BF_{10} > 100$; r = Pearson's r; SP = Statistical Power at $\alpha < .05$;

Table 5. Bayesian correlations between the VR-EAL and the paper-and-pencil tests

Paper-and-Pencil Scores	VR-EAL Scores	r	BF_{10}	SP

CAMPROMPT – Total	Total PM	0.82***	3.20e+9	~ 100%
CAMPROMPT - Event Based	Event Based PM	0.73***	3.97e+3	~ 100%
CAMPROMPT - Time Based	Time Based PM	0.67***	2.61e+2	~ 100%
RBMT – Immediate Recall	Immediate Recognition	0.77***	7.34e+7	~ 100%
RBMT – Delayed Recall	Delayed Recognition	0.82***	3.90e+9	~ 100%
TEA – Map Total Score	Selective Visual Attention Accuracy	0.48**	50.53	95%
TEA – Map Total Score	Selective Visual Attention Speed	0.46**	34.99	93%
RSAT – Accuracy	Selective Visual Attention Accuracy	0.43*	16.94	89%
RSAT – Accuracy	Selective Visuospatial Attention Total Score	0.61***	2101	99%
RSAT – Speed	Selective Visuospatial Attention Speed	0.49**	63.15	96%
RSAT – Accuracy	Selective Visuospatial Attention Accuracy	0.58***	778.50	99%
TEA -Elevator Counting with Distraction	Selective Auditory Attention	0.70***	8.91e+4	~ 100%
BADS – Key Search	Planning	0.80***	4.65e+8	~ 100%
CTT – 1	Planning	0.47**	41.74	94%
CTT – 2	Planning	0.51***	109.73	97%
CTT – 1	Cooking Task	0.70***	9.88e+4	~ 100%
CTT – 2	Cooking Task	0.80***	8.75e+8	~ 100%
BADS – Key Search	Cooking Task	0.62***	2.99e+3	99%

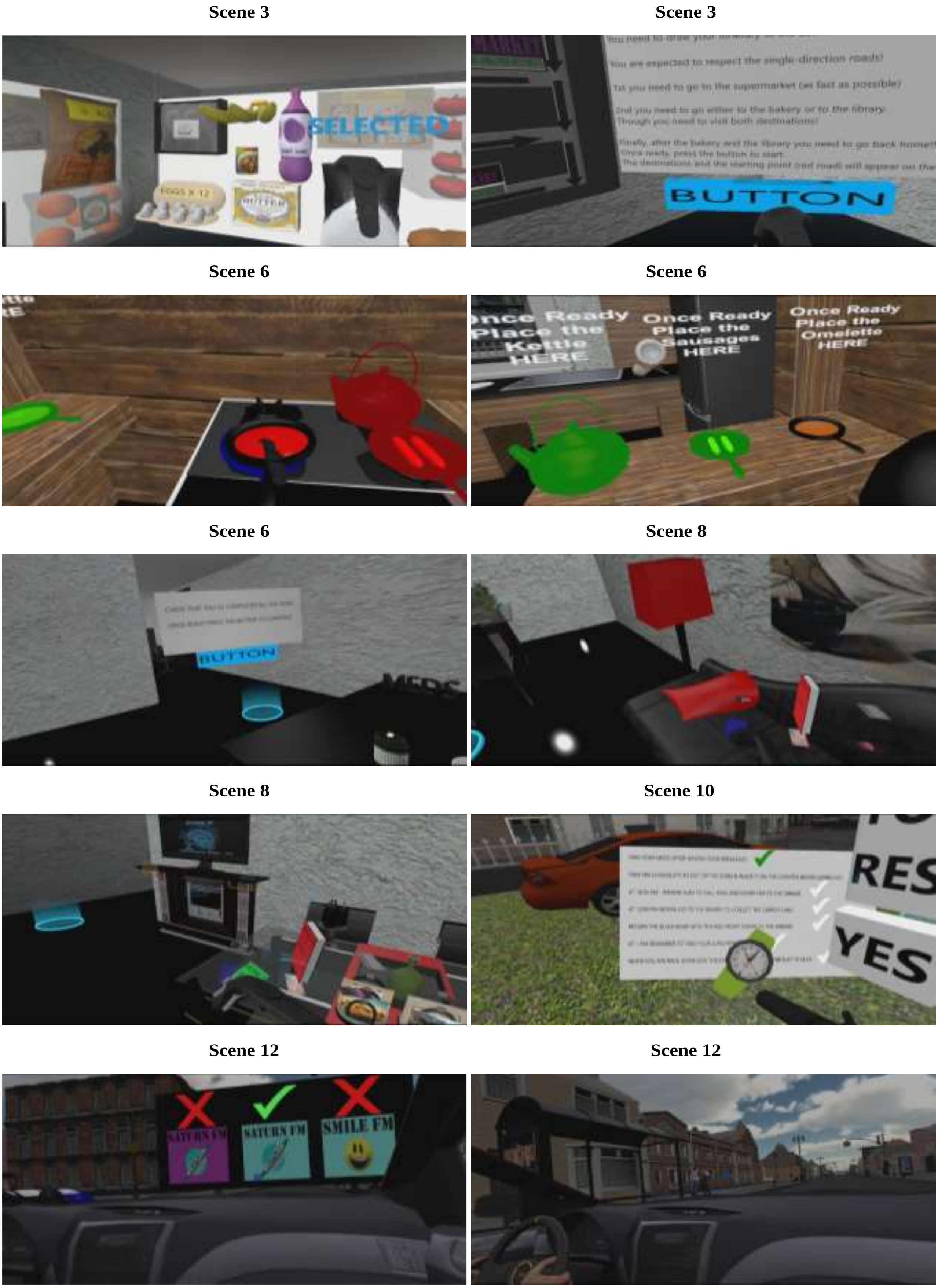
The alternative hypothesis specifies that the correlation is positive. * $BF_{10} > 10$; ** $BF_{10} > 30$; *** $BF_{10} > 100$; r = Pearson's r; SP = Statistical Power at $\alpha < .05$; VR-EAL = Virtual Reality Everyday Assessment Lab; CAMPROMPT = Cambridge Prospective Memory Test; RBMT = Rivermead Behavioral Memory Test; TEA = Test of Everyday Attention; RSAT = Ruff 2 and 7 Selective Attention Test; BADS = Behavioral Assessment of the Dysexecutive Syndrome; CTT = Color Trails Test

Table 6. Comparison between administration time and participants' ratings of verisimilitude and enjoyment for the VR-EAL and paper-and-pencil tests

Paper-and-Pencil Test		**VR-EAL**	**BF_{10}**	**SP**
Total Administration Time	>	VR-Session Time	1.224e+11***	~ 100%
Testing Pleasantness	<	VR-Testing Pleasantness	188,842***	~ 100%
Total Verisimilitude	<	VR-EAL Verisimilitude	4.898e +15***	~ 100%
Ecologically Valid Tests/Tasks Verisimilitude	<	VR-EAL Verisimilitude	3.575e +13***	~ 100%
CAMPROMPT Verisimilitude	<	PM Verisimilitude	1.179e +9***	~ 100%
BADS Key Search Verisimilitude	<	Planning Verisimilitude	1.950e +13***	~ 100%
CTT Verisimilitude	<	Cooking Task Verisimilitude	6.849e +21***	~ 100%
RSAT Verisimilitude	<	Visuospatial Attention Verisimilitude	2.635e +13***	~ 100%
TEA Map Verisimilitude	<	Visual Attention Verisimilitude	3.774e +12***	~ 100%
TEA Elevator Counting with Distraction Verisimilitude	<	Auditory Attention Verisimilitude	4.36e +11***	~ 100%
RBMT Story Recall Verisimilitude	<	Episodic Memory Verisimilitude	1.244e +7***	~ 100%

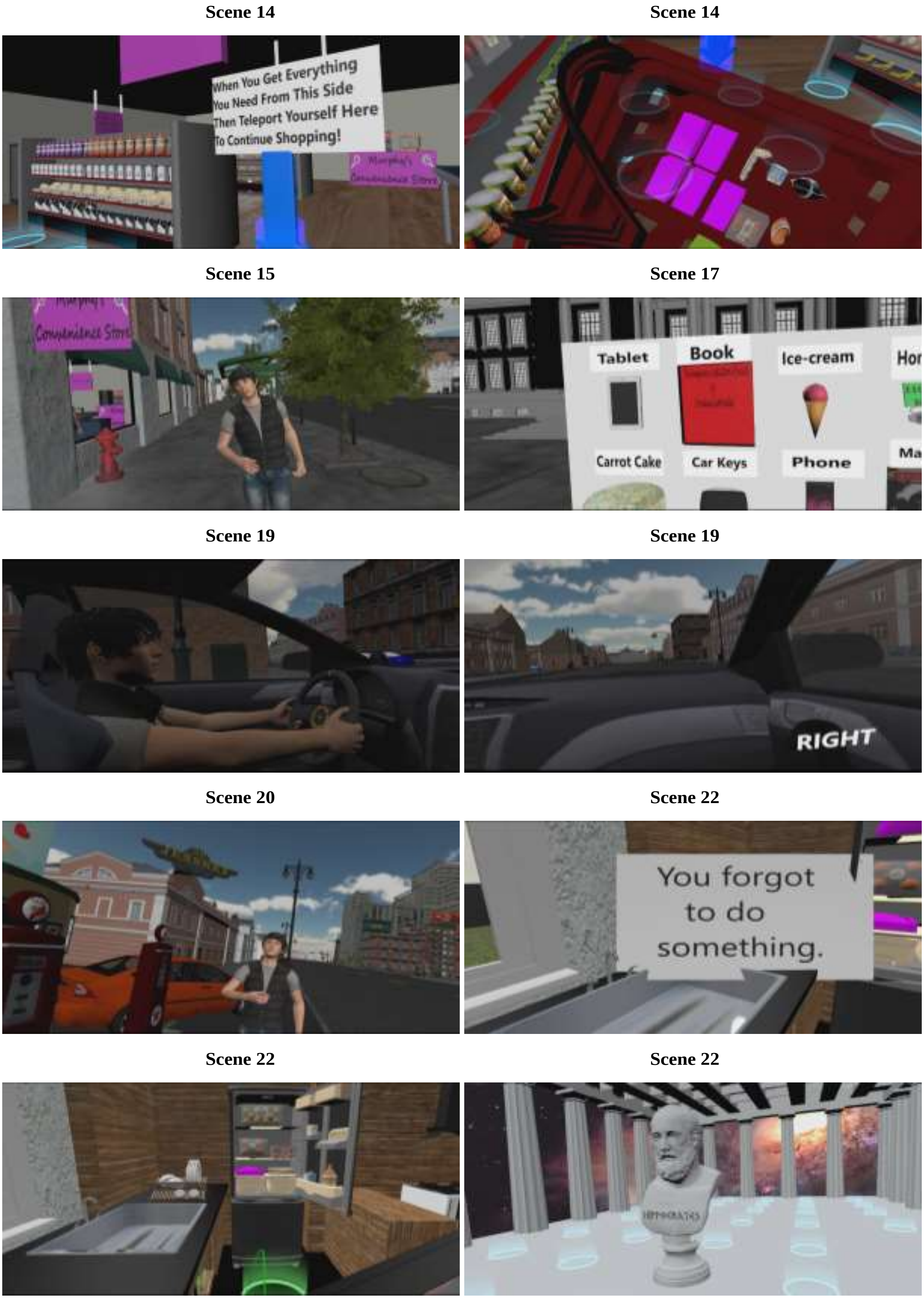
* $BF_{10} > 10$; ** $BF_{10} > 30$; *** $BF_{10} > 100$; SP = Statistical Power at $\alpha < .05$; VR-EAL = Virtual Reality Everyday Assessment Lab; CAMPROMPT = Cambridge Prospective Memory Test; BADS = Behavioral Assessment of the Dysexecutive Syndrome; CTT = Color Trails Test; RSAT = Ruff 2 and 7 Selective Attention Test; TEA = Test of Everyday Attention; RBMT = Rivermead Behavioral Memory Test; PM = Prospective Memory; VR = Virtual Reality

Figure 1. VR-EAL Storyline: Scenes 3 - 12.



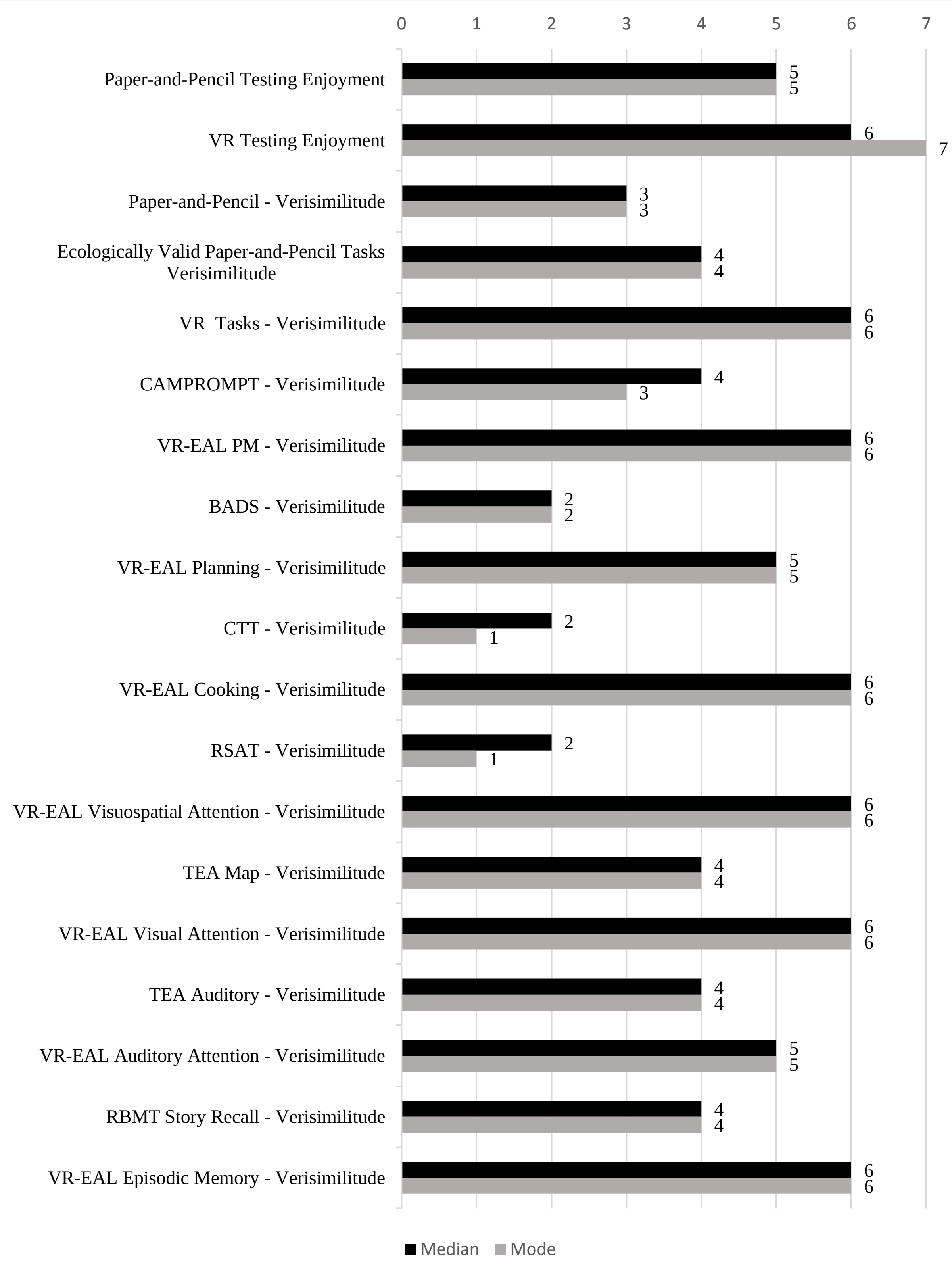
Derived from Kourtesis et al., (2020).

Figure 2. VR-EAL Storyline: Scenes 14 – 22



Derived from Kourtesis et al., (2020).

Figure 3. Self-report verisimilitude and enjoyment of the VR-EAL and paper-and-pencil tests.



VR = Virtual Reality; CAMPROMPT = Cambridge Prospective Memory Test; PM = Prospective Memory; VR-EAL = Virtual Reality Everyday Assessment Lab; BADS = Behavioral Assessment of the Dysexecutive Syndrome; CTT = Color Trails Test; RSAT = Ruff 2 and 7 Selective Attention Test; TEA = Test of Everyday Attention; RBMT = Rivermead Behavioral Memory Test