frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 26 May 2023
DOI 10.3389/frvir.2023.1138240

OPEN ACCESS

EDITED BY
Kiyoshi Kiyokawa,
Nara Institute of Science and Technology (NAIST), Japan

REVIEWED BY
Panagiotis Kourtesis,
National and Kapodistrian University of Athens, Greece
Adnan Fateh,
University of Central Punjab, Pakistan

*CORRESPONDENCE
Erik Seesjärvi,
erik.seesjarvi@helsinki.fi

RECEIVED 05 January 2023
ACCEPTED 25 April 2023
PUBLISHED 26 May 2023

CITATION
Seesjärvi E, Laine M, Kasteenpohja K and Salmi J (2023), Assessing goal-directed behavior in virtual reality with the neuropsychological task EPELI: children prefer head-mounted display but flat screen provides a viable performance measure for remote testing.
Front. Virtual Real. 4:1138240.
doi: 10.3389/frvir.2023.1138240

Assessing goal-directed behavior in virtual reality with the neuropsychological task EPELI: children prefer head-mounted display but flat screen provides a viable performance measure for remote testing

Erik Seesjärvi[1,2]*, Matti Laine[3], Kaisla Kasteenpohja[1] and Juha Salmi[4,5,6]

[1]Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland, [2]Child Neurology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, [3]Department of Psychology, Åbo Akademi University, Turku, Finland, [4]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland, [5]MAGICS, Aalto University, Espoo, Finland, [6]Aalto Behavioral Laboratory, AMI-centre, Aalto University, Espoo, Finland

Background and objective: EPELI (Executive Performance of Everyday LIving) is a Virtual Reality (VR) task that was developed to study goal-directed behavior in everyday life contexts in children. In this study, we had 72 typically developing 9- to 13-year-old children to play EPELI with an immersive version implemented with a head-mounted display (HMD) and a non-immersive version employing a flat screen display (FSD) in a counterbalanced order to see if the two versions yield similar results. The children's everyday executive functions were assessed with the parent-rated Behavior Rating Inventory for Executive Functions (BRIEF) questionnaire. To assess the applicability of EPELI for online testing, half of the flat screen display version gameplays were conducted remotely and the rest in the laboratory.

Results: All EPELI performance measures were correlated across the versions. The children's performance was mostly similar in the two versions, but small effects reflecting higher performance in FSD-EPELI were found in the measures of Total score, Task efficacy, and Time-based prospective memory score. The children engaged in more active time monitoring in FSD-EPELI. While the children evaluated the feeling of presence and usability of both versions favorably, most children preferred HMD-EPELI, and evaluated its environment to be more involving and realistic. Both versions showed only negligible problems with the interface quality. No differences in task performance or subjective evaluations were found between the home-based and laboratory-based assessments of FSD-EPELI. In both EPELI versions, the efficacy measures were correlated with BRIEF on the first assessment, but not on the second. This raises questions about the stability of the associations reported between executive function tasks and questionnaires.

Conclusions: Both the HMD and FSD versions of EPELI are viable tools for the naturalistic assessment of goal-directed behavior in children. While the HMD

version provides a more immersive user experience and naturalistic movement tracking, the FSD version can maximize scalability, reachability, and cost efficacy, as it can be used with common hardware and remotely. Taken together, the findings highlight similarities between the HMD and FSD versions of a cognitively complex VR task, but also underline the specific advantages of these common presentation modes.

KEYWORDS

prospective memory, ecological validity, executive functions, online testing, naturalistic task, serious gaming

1 Introduction

The literature of virtual reality (VR) based cognition research is expanding at a rapid pace, reflecting the increasing availability of VR systems and their technological advancements (Cipresso et al., 2018; Krohn et al., 2020). Notably, VR has been suggested as an ideal way of implementing new naturalistic paradigms that mimic real-life functions and situations (Chan et al., 2008; Parsons, 2015; Parsons et al., 2017), as it offers safe and flexible ways to create various easily-reproducible environments and allows diverse behavioral responses (e.g., movements of the eyes, head, and body) to be measured accurately (see Campbell et al., 2009). Such naturalistic tasks could complement more traditional cognitive laboratory tasks which are often repetitive, contain a limited set of stimuli, and permit only restricted behavioral responses such as a button press (Hatfield, 2002). Naturalistic tasks can also be more sensitive to cognitive impairments in situations where more traditional tasks fail to detect them (Shallice & Burgess, 1991; Cipresso et al., 2014) and could offer better predictive value for everyday functioning (e.g., Burgess et al., 2006; Chan et al., 2008; Parsons et al., 2017; Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). Different VR environments permit the researcher to present dynamic stimuli in a way that allows for both the veridical control of laboratory measures and the verisimilitude of naturalistic observation of real-life situations (Parsons, 2015).

VR can be accomplished through several technical solutions that differ, among other things, in their immersiveness. An immersive VR system can be defined as one that allows the participant to perceive the environment and interact with it through natural sensorimotor contingencies (Slater & Sanchez-Vives, 2016), or as a system that blurs the lines between the physical and virtual worlds (Suh & Prophet, 2018). High immersiveness requires effective sensory substitution, which depends on factors like wide field-of-view, stereo vision/sound, head tracking for changing the field of view, short latency from head move to display, and high-resolution displays (Slater & Sanchez-Vives, 2016). In broad terms, the systems implemented with head-mounted displays (HMDs) and dedicated position-tracking controllers or camera-based hand tracking can be regarded as immersive VR, and the systems based on flat screen displays (FSDs) and more traditional interaction devices (e.g., keyboards, joysticks, and mice) as non-immersive VR (e.g., Suh & Prophet, 2018; Di Natale et al., 2020). The sense of presence is a subjective correlate of immersion and can be defined as having the illusion of "being there" in the VR environment while being aware about not actually being there (Slater & Sanchez-Vives, 2016). Importantly, the sense of presence can be considered to be a key aspect of a virtual experience and its ecological validity, as it can be argued that only when the participant is having a strong sense of presence in the virtual experience, s/he will show same kind of reactions to it that may be expected under real-life circumstances (Kothgassner & Felnhofer, 2020) and perform the tasks as s/he would do them in real life (Pan & Hamilton, 2018; Slater, 2018).

HMDs and the related peripherals have several benefits over traditional FSDs and their interaction devices. They can more closely emulate real-life sensory-motor experiences than FSDs by matching the criteria for high immersiveness to a greater extent. For example, turning the head with an HMD alters the view in the virtual world in parallel with the actual physical movements, which cannot be accomplished with common FSDs. Typical hand controllers of the current HMD systems track their rotation and position, so turning and moving the physical controller leads to similar rotations and movements in the controller projected to the virtual space. HMDs offer a stereoscopic visual experience, and with current hardware the field of view (FOV) is markedly larger than that of a typical FSD (Parsons, 2015). Furthermore, HMDs usually block the view of the surrounding physical environment completely, which can further increase immersiveness (see Slater, 2018). These differences can lead to higher perceived presence when using HMDs (Tan et al., 2015; Pallavicini et al., 2018; Makransky et al., 2019; Pallavicini et al., 2019; Pallavicini & Pepe, 2019; Yao & Kim, 2019; Chang et al., 2020; Li et al., 2020; Caroux, 2023) and have behavioral implications, such as greater physical effort with HMDs (Yao & Kim, 2019).

There are also potential disadvantages with HMDs when compared to FSDs. To avoid some of these disadvantages, the implementation of HMD-based neuropsychological tasks calls for special consideration for aspects such as how controls that facilitate naturalistic interactions are achieved, how these controls are learnt by everyone so that gamers will not have an advantage over those participants who do not play regularly or at all, what kind of hardware is required for smooth graphics, how the measurement of targeted cognitive domains or behavior is accomplished, and importantly, how potential cybersickness symptoms like nausea, dizziness, and headache are avoided (Kourtesis et al., 2020). The earlier HMDs were sometimes reported to cause cybersickness symptoms (Bohil et al., 2011), but these have been markedly reduced or have disappeared with the newer generation of HMDs (Kourtesis et al., 2019; Weech et al., 2019). Eradicating cybersickness is not vital only for the comfort of the participant but also for ecological validity, as the sense of presence and cybersickness are negatively associated (Weech et al., 2019). Recent studies have provided insights on how cybersickness is related to display lag

in virtual and physical head pose (Palmisano et al., 2020) and how it can be countered by dynamic FOV restriction (Teixeira & Palmisano, 2021). This information helps researchers to design their paradigms in a way that minimizes the risk of these adverse effects. Still, cybersickness symptoms might arise in situations where there is a conflict between perceived and physical movements (Bohil et al., 2011; Palmisano et al., 2020), and some individuals, such as those with autism spectrum disorder, might be especially prone to them (Parsons et al., 2017). Because of these potential adverse effects, FSDs might be the preferred choice in some situations, for example in wheelchair training (Alapakkam Govindarajan et al., 2022) or in a race driving simulation (Walch et al., 2017). FSDs are widely available, and the related interfaces and operating systems are highly familiar even for less technically oriented users. Using HMD systems might require additional investment, and the users may sometimes need training to use the interfaces and software. Overall, the FSD systems are cost-efficient, easy-to-use, and flexible, especially in certain situations such as remote testing with automated web platforms.

As both HMD- and FSD-based systems provide means for implementing similar tasks but have different advantages as discussed above, it is essential to compare their unique characteristics so that informed decisions can be made when choosing between the two. Making such decisions for naturalistic neuropsychological tasks is currently hampered by the small number of studies that compare the two technologies by implementing such tasks in both. Furthermore, because of the rapid advances in the HMD technology, the results of earlier studies with older HMD models might not apply to the current hardware.

Regarding learning outcomes, some studies have compared HMDs and FSDs with a task that was implemented similarly between the conditions (e.g., Makransky et al., 2019; Ventura et al., 2019; Barrett et al., 2022). In a within-subjects study comparing learning in a science lab simulation in HMD and FSD conditions, Makransky et al. (2019) found that students reported having a stronger sense of presence during the HMD condition, but they also learned less and had significantly higher cognitive load based on electroencephalogram (EEG). Studying category learning, Barrett et al. (2022) found no significant group differences in learning accuracy between HMD and comparison conditions (FSD with 3D and 2D stimuli), although the participants in the HMD group had increased fixation counts. Contrasting these findings, Ventura et al. (2019) found stronger memory performance after immersive HMD condition than non-immersive tablet flat screen condition. Thus, the use of either HMDs or FSDs may result in better memory performance and more effective learning, but this could also depend on the specific task and hardware.

Several FSD-based traditional cognitive tasks, which include only a small set of stimuli and behavioral responses, have also been successfully adopted to HMDs. In their original form, many of these laboratory tasks have limitations such as their two-dimensional environment, non-naturalistic responses (e.g., using a keyboard or response box) and stimulus dynamics, and a substantial divergence from looking realistic (Kourtesis & MacPherson, 2021), which affects their immersiveness. Although the original versions of these tasks have low immersiveness, some of their HMD adaptations have taken use of the immersive capabilities of the technology. As an example, Armstrong et al. (2013) compared a Stroop task embedded in an HMD-VR scene with a FSD version and paper-and-pencil version of the same task. They found the reaction time measures in all three conditions (Word reading, Color naming, and Interference) to be correlated between the VR and the FSD version (r = 0.64–0.75), but between VR Stroop Task and the paper-and-pencil version reaction time was only correlated in the Interference condition (r = 0.49). Another cognitive task for which several different HMD versions exist is the Continuous Performance Test. However, several of them do not merely aim to be faithful replications of the FSD versions, but also take advantage of HMDs' extended possibilities, for example, by including extraneous distractors (see the meta-analysis by Parsons et al., 2019). To study the convergent validity of an HMD-based Continuous Performance Test coined as AULA, Nesplora, Díaz-Orueta et al. (2014) compared it to a FSD version (Conners' Continuous Performance Test) in a group of children aged 6–16 years. They found that all key measures (omissions, commissions, reaction time, reaction time variability) were correlated between the two versions (ρ = 0.36–0.79). Li et al. (2020) implemented FSD and HMD versions of the Posner task in a within-subject design and studied the related attentional processes, which were found to be enhanced in the HMD version, according to both behavioral data and EEG responses. Based on these findings, the authors suggested that the allocation of attentional resources would be more effective with an HMD compared to the FSD condition. Moreover, their participants evaluated that the sense of presence was strong during the HMD but weak during the FSD condition. In sum, these studies suggest that HMD and FSD versions of these cognitive tasks seem to be measuring the same phenomena, but differences between the two platforms can affect participants' performance and subjective experience to some extent.

Another important methodological issue concerns the pros and cons of laboratory-based versus home-based testing via the Internet. Home-based remote testing can be an attractive and efficient option in many research and clinical settings, such as in a large-scale data collection (Feenstra et al., 2017). It is especially well-suited for the FSD systems, as the required hardware (i.e., regular home computers) are widely available. As the COVID-19 pandemic has shown, face-to-face testing might become impossible for reasons that are beyond researcher's control (Zuber et al., 2021). However, it is not guaranteed that unsupervised remote testing with varying hardware would produce results as reliable as those from laboratory-based testing with fixed equipment. While some authors have found comparable performance between web- and laboratory-based testing (Germine et al., 2012), others have found some disparity between laboratory and online results (Crump et al., 2013). Backx and others (2020) used a within-subject design to examine the comparability of performance in the Cambridge Neuropsychological Test Automated Battery (CANTAB) under two conditions: an unsupervised web-based test situation and a typical in-person lab-based assessment. The test-retest stability was found to be comparable to previous studies with CANTAB, as the intraclass correlations ranged from 0.23 to 0.67, with high correlations (>0.60) in 3/9 performance indices and 2/5 reaction time measures. Performance indices did not differ between the

conditions and generally showed satisfactory agreement, and learning effects were present in 3/9 indices. However, reaction times were slower during web-based assessments, which undermined their equivalence and agreement. This was likely due to variations in computer hardware. Also using a within-subjects design, Zuber and others (2021) found moderate-to-high correlations (r = 0.56–0.68) between laboratory and online assessment in a prospective memory task called the Geneva Space Cruiser. Overall, while remote testing is an attractive option for various research and clinical settings, each online implementation needs to be studied separately to ensure its applicability and the robustness of the results.

There are several studies on naturalistic VR tasks that simulate daily functions and activities. Some have used FSDs (e.g., Rand et al., 2009; Jovanoski et al., 2012; Raspelli et al., 2012; Cipresso et al., 2014; Ruse et al., 2014) while others have employed HMDs (e.g., Barnett et al., 2021; Chicchi Giglioli et al., 2021; Kourtesis et al., 2021; Ouellet et al., 2018; Parsons & Barnett, 2017; Porffy et al., 2022; see also the reviews by Neguț et al., 2016; Parsons, 2015; and Pieri et al., 2023). Regarding the Multiple Errands Test that was at first devised to be performed in real-life environments (Shallice & Burgess, 1991; see also Rotenberg et al., 2020), there are several desktop FSD versions have been implemented (Rand et al., 2009; Jovanoski et al., 2012; Raspelli et al., 2012; Cipresso et al., 2014), as well as a simplified tablet version to serve as a brief screening tool (Webb et al., 2021). These studies have not included any direct comparison between FSDs and HMDs, although the tasks included could be implemented with small adjustments in both. However, some other studies with relatively recent HMD hardware have compared the two technologies directly by implementing the same task with both, albeit their scenarios were not taken directly from ordinary daily life (Brooks et al., 2017; Chang et al., 2020). Brooks and others (2017) compared the HMD and FSD versions of a military flight simulator in a within-subjects study and found no difference in target detection performance between the two versions, but their participants reported higher mental workload and discomfort when using the HMD. Contrasting these findings, Chang and others (2020) performed a between-subjects study using a driving simulation with an embedded Stroop task to compare HMD and FSD conditions. They found that participants using an HMD performed better for the virtual driving but did not differ in self-reported mental effort and psychophysiological responses compared to the FSD condition. However, the authors found that users in the FSD condition had a shorter average reaction time on the Stroop trials, which they interpreted as an indication that driving required more selective attention in the HMD condition. This may have led to slower responses in the Stroop task. These two studies as well as the before-mentioned studies of learning outcomes and traditional cognitive tasks provide an important reference for further studies comparing FSD and HMD platforms but leave open what differences could exist between the FSD and HMD versions of tests with more open-ended naturalistic scenarios.

Recently, we developed EPELI (Executive Performance in Everyday LIving) with HMD to study goal-directed behavior of children in everyday contexts (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). To our knowledge, EPELI is the first naturalistic VR task for children that requires the participants to carry out multiple tasks from memory by navigating a virtual home and interacting with the relevant target objects, while keeping track of the time and ignoring non-relevant distracting objects and events. Successful performance in such goal-directed actions requires attentional, executive, and memory resources (Seesjärvi et al., 2022a). We have previously shown that the most important measures in HMD-EPELI show acceptable internal consistency, and the measure of task efficacy in particular is associated with parent-rated problems of executive function (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). The children evaluated HMD-EPELI to be very enjoyable and reported only negligible cybersickness symptoms (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). In a study using HMD-EPELI and a sample of school-aged children, some measures were associated with age (older children outperforming younger), gender (girls outperforming boys), and verbal encoding ability (children with better ability outperforming those with worse; Seesjärvi et al., 2022a). Notably, there were no significant associations of gaming background, task familiarity, or HMD type (Oculus GO vs. Pico Neo 2 Eye) with the EPELI measures (Seesjärvi et al., 2022a). Even though HMD-EPELI does take advantage of additional benefits of HMDs, such as using natural head movements for looking around the VR environment, the task itself is also well-suited for the FSD systems.

The main aim of the current study was to compare HMD and FSD implementations of a naturalistic VR task, EPELI. To our knowledge, this is the first study to make such a comparison with a naturalistic task that calls for goal-directed behavior in varied but typical everyday scenarios. Therefore, the study was expected to make a valuable contribution to the VR-based literature of cognition research, as these function-led paradigms take full advantage of the new technological possibilities and can be the hallmark of VR-based cognition research (Parsons et al., 2017). For this study, we developed a FSD version of EPELI that enabled us to examine the similarities and differences between the HMD and FSD implementations in a counterbalanced within-subjects design. A successful FSD implementation of an HMD-based naturalistic cognitive task could significantly widen its applicability in various situations, especially in remote testing. Therefore, we also studied the feasibility of parent-supervised remote testing by asking half of the participants to perform FSD-EPELI at home. Furthermore, we wanted to re-examine the associations between EPELI efficacy measures and parent-rated difficulties in executive function, which have previously been reported between HMD-EPELI and BRIEF (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). Finally, inter-version (FSD/HMD) and test-retest correlations were analyzed, as these provide important insights into the reliability and stability of a task.

The specific research aims were as follows:

1) To examine similarities and differences in task performance measures between the FSD and HMD versions and learning effects between the first and second assessment.
2) To probe similarities and differences in subjective experience ratings between FSD- and HMD-EPELI.
3) To study similarities and differences in FSD-EPELI task performance measures and subjective experience ratings between experimenter-supervised laboratory testing and parent-supervised home testing.
4) To inspect possible associations between FSD- and HMD-EPELI efficacy measures and parent-rated EF difficulties (BRIEF questionnaire; Gioia et al., 2000).

5) To assess the inter-version (FSD vs. HMD) correlations and test-retest stability of EPELI.

2 Materials and methods

2.1 Participants

The study included 101 typically developing children from Kirkkonummi and Espoo, Finland (see Supplementary Material for further information about the recruitment process). The inclusion criteria were a) native language Finnish and b) age of 9–12 years when recruited for the study. The exclusion criteria were a) any psychiatric, behavioral, or neurodevelopmental disorders (F00–F99 in ICD-10; World Health Organization, 1992) and b) decision of special support at school. For 29 children, the EPELI data for one of the two sessions (see 2.3 Procedure) was missing because of dropping out of the study after the first session or due to technical problems. Thus, the final sample comprised 72 typically developing children (29 girls and 43 boys, mean age of all participants 11.0 years and age range 9.4–13.0 years; for descriptive statistics, see Supplementary Table S1), who had successfully taken part in both sessions. The study was approved by the Helsinki University Hospital Ethics Committee, and informed consent according to the Declaration of Helsinki was obtained from children and their parents. Each child received four movie tickets for participating.

2.2 The EPELI task

EPELI (https://aalto.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=3eb4836f-1238-4f27-853a-ad3700745b31; for the original description, see Seesjärvi et al., 2022b) is a naturalistic task of goal-directed behavior. It was designed with equal contributions by ML, JS, and ES, inspired by tasks simulating everyday life requirements, such as the Virtual Week (Rendell & Craik, 2000) and Multiple Errands Test (Shallice & Burgess, 1991). With all 13 scenarios and the practice session, EPELI takes on average approximately 27 min to complete. It was first implemented with HMD technology and then converted to FSD for this study. The key differences between the versions are as follows: a) in the FSD version, the participant uses a mouse/trackpad to change the direction of the view, whereas when using HMD, this can be accomplished by rotating the head; b) in the FSD version, the FOV is markedly smaller (101 versus approximately 25–60°, see Supplementary Material); c) in the HMD version, the view of the surrounding physical environment is blocked by the goggles, and the technology provides stereoscopic view; d) while in both versions participants interact with objects by pointing at them and clicking a button, in the HMD version this can be done independently from the direction of the view by rotating the hand controller until the ray coming from the virtual hand controller object is pointing at the desired object, whereas in the FSD version the participant is required to turn the direction of the view until the desired object is located in the crosshairs in the middle of the screen (see Figure 1D); e) In the HMD version, the clock is viewed by raising the hand and looking at the virtual hand controller object (see Figure 1C). In the FSD version, there is a white circle at the lower right corner of the screen that reveals a clock when the second mouse/trackpad button is pressed (see Figure 1D). In the HMD version, the participants used Oculus Go goggles (2560 × 1440 resolution, 60/72 Hz refresh rate, 16:9 aspect ratio, and 101-degree horizontal FOV) and the related hand controller in a sitting position (see Figure 1A). In the FSD version, the participants used typical laptop/desktop computers and a web browser (see Figure 1B). For further details on the version differences, see Supplementary Material.

For both versions, the eight EPELI performance measures (Total score, Task efficacy, Navigation efficacy, Controller motion, Total actions, Time-based prospective memory score = TBPM, Clock checks, and Event-based prospective memory score = EBPM) described in an earlier study (Seesjärvi et al., 2022a) were included in the analyses. The only difference in the descriptions concern the measure of Controller motion in FSD-EPELI. In the FSD version, rotating the view needs to be done with mouse/trackpad as opposed to the natural head movement utilized in the HMD version. Therefore, this measure is likely to tap somewhat different aspects of behavior in the two versions.

2.3 Procedure

The study included two assessment sessions, one with HMD-EPELI and the other with FSD-EPELI, performed in a counter-balanced order 3.3–10.5 months apart (Figure 2). After both EPELI versions, the children orally answered a translated version of the Simulator Sickness Questionnaire (Kennedy et al., 1993; see also Seesjärvi et al., 2022a) and a shortened version of the Presence Questionnaire 3.0 (Witmer et al., 2005; see also Seesjärvi et al., 2022a). After HMD-EPELI, they also answered a gaming experience questionnaire (Seesjärvi et al., 2022b). To probe their familiarity with the task contents, the children were also asked *"From a scale of 1 (not at all) to 7 (very much), how much have you performed similar tasks in real life?"*. After FSD-EPELI performed at home, the family also filled out a hardware questionnaire (see Supplementary Table S2). The parents filled out the Behavior Rating Inventory for Executive Functions questionnaire (BRIEF; Gioia et al., 2000), from which the raw score of Global Executive Composite (GEC) was used. There was no difference in the average time between the sessions between the groups who performed the HMD part and the FSD part first, but for both groups, the delay between the sessions was longer than planned, as affected by the restrictions imposed by the global COVID-19 pandemic. All participants performed the HMD-EPELI session in the laboratory, while the FSD-EPELI session was performed in laboratory or an equivalent dedicated school room by 37 children and at home by 35 children. The children were assisted and supervised in laboratory by one of the researchers or by a trained research assistant, and at home by a parent. After performing EPELI (either HMD or FSD) and the related questionnaires in the second session, the children were also asked which version was more realistic, preferable, and easier to play, with response alternatives HMD/FSD/"I don't know". For detailed information about the procedure, see Supplementary Material.

2.4 Statistical analyses

All statistical analyses and data visualization were done in R version 4.0.3 (R Core Team, 2020) with the additional packages data.table (Dowle & Srinivasan, 2021), stringr (Wickham, 2019), stringi

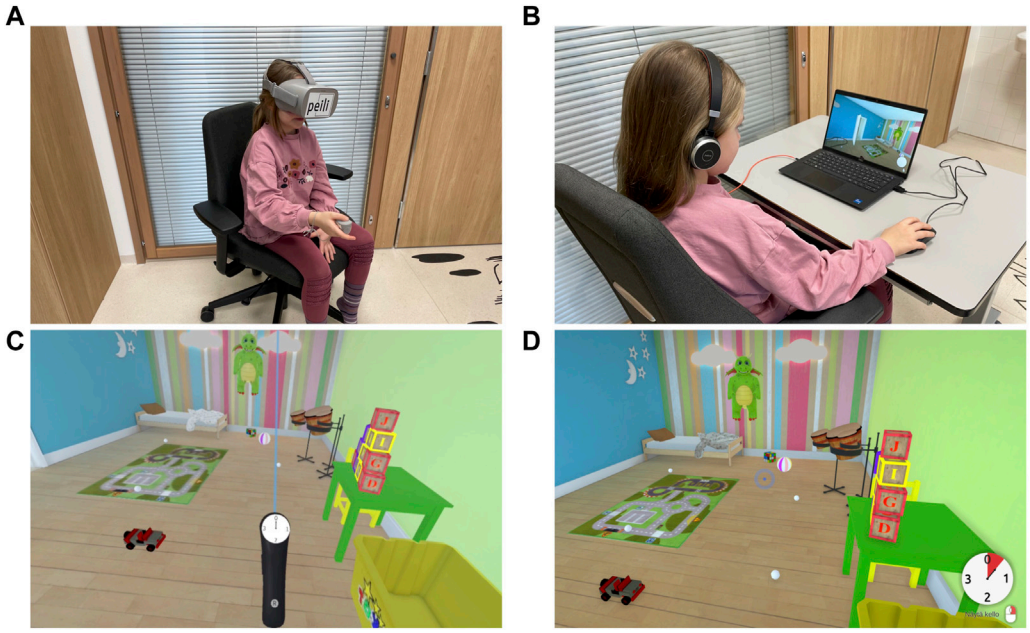

FIGURE 1
Pictures and screenshots from an EPELI session. **(A)**, a participant performing HMD-EPELI. **(B)**, the same participant during FSD-EPELI. **(C)**, a screenshot from HMD-EPELI showing the virtual hand controller with the clock. **(D)**, a screenshot from FSD-EPELI showing the clock in the lower right corner of the screen and the crosshairs in the middle of the screen.

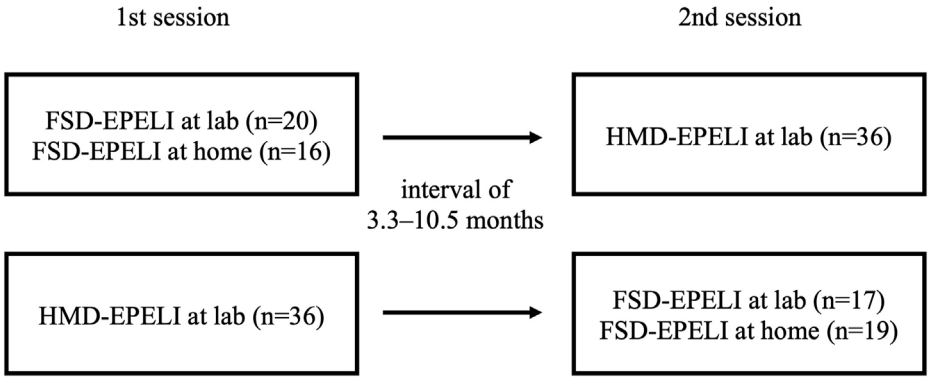

FIGURE 2
The study design.

(Gagolewski, 2020), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), effectsize (Ben-Shachar et al., 2020), tidyverse (Wickham et al., 2019), ppcor (Kim, 2015), dplyr (Wickham et al., 2021), ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), patchwork (Pedersen, 2020), and psych (Revelle, 2020).

First, the data were inspected for missing values, data handling errors, and possible outliers. The questionnaires to be filled after FSD-EPELI (see 2.3 Procedure) were missing from six participants in the home group. Also, one parent had not answered the BRIEF questionnaire after FSD-EPELI. Univariate outliers in EPELI, BRIEF and presence questionnaire were first identified visually and confirmed numerically using a cutoff of three standard deviations above or below the mean. For FSD-EPELI, this was done separately for the lab and home groups. As a result, three HMD-EPELI gameplays, two FSD-EPELI gameplays and two BRIEF questionnaires were removed from the data, as at least one variable was confirmed to be an outlier. The observations removed comprised 3.2 % of the total data. The data was then checked for multivariate outliers using the same cutoff, but none were found. The average administration time was equal between the versions (on average 27.5 min for the FSD version and 27.8 min for the HMD version, $t(122) = -0.59$, $p = 0.55$) and very close to what had been observed in previous studies (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b).

Similarities and differences in task performance between FSD- and HMD-EPELI and between the first and second sessions were evaluated with general linear mixed models (LMM) with each EPELI variable except Controller motion at time as the dependent variable, EPELI version (FSD/HMD) and time (first/second session) as fixed factors, and participant as a random factor. This analysis was not performed for Controller motion, as it measures somewhat different aspects in the FSD and HMD versions due to the differences in the control interfaces. In the models with Total actions and Clock checks as the dependent variable, the error terms distributions did not follow normal distribution. Therefore additional generalized LMMs using Poisson distribution were fitted for these variables. These models yielded very similar results, and therefore only the general LMMs are reported. The lmer function from the lme4 package was used for the LMMs, and the effect sizes were estimated with t_to_d function from the effectsize package. Effect sizes were estimated as Cohen's *d* and interpreted as suggested by Conner et al., 2022 as small (>0.20), medium (>0.50), or large (>0.80).

Similarities and differences in subjective experience between the FSD and HMD versions were assessed as follows. First, LMMs with each Presence questionnaire item as the dependent variable while using version (FSD/HMD) and time (first/second session) as fixed factors and participant as a random factor. As the error term distributions in the models of questions 5, 6, 7, 8, and 12 were not normal, the main effects of version and time on these questions were confirmed with Wilcoxon signed-rank tests with continuity correction. These results are in line with those obtained with LMMs and are not shown. Second, the difference in Simulator Sickness Questionnaire between the version was tested with Wilcoxon signed rank test with continuity correction. Third, any possible differences in the three questions regarding head-to-head comparison of the versions (FSD/HMD) were tested with exact binomial tests.

Similarities and differences in FSD-EPELI performance and subjective experiences between laboratory and home testing were assessed with LMMs using each EPELI measure and Presence questionnaire item at a time as the dependent variable, place of the assessment (lab/home) and time (first/second session) as fixed factors, and participant as a random factor.

The associations between EPELI efficacy measures and BRIEF were examined with bivariate correlations. Based on visual inspection, all distributions were near to normal and thus Pearson's correlation coefficients were used. The correlations were calculated both for each EPELI version (HMD/FSD) and its corresponding BRIEF questionnaire, and for each EPELI session (first/second) and its corresponding BRIEF questionnaire.

The inter-version (FSD/HMD) and test-retest stabilities of EPELI were first assessed with bivariate correlation coefficients to allow the comparison with earlier literature. Then, intraclass correlations (ICCs) were calculated with single-rating, absolute agreement, two-way random effect models (ICC 2,1 in Martel et al., 2015), to account not only for the within-subject change but also for the differences in the group means between the versions. For ICCs, function ICC from package psych was used. To assess the effect of one factor (version or time) while controlling for the other but without accounting for the within-subject variation, partial correlations were also provided for both inter-version and test-retest correlations with the other factor as a covariate. The partial correlations were calculated with function pcor from package ppcor and were chosen as the primary correlation measures. All distributions in both EPELI versions were visually evaluated to be near to normal and Pearson's correlation coefficients were used, except those of Total actions, which were strongly skewed to the right. To evaluate the effect of this skewness to the results, these distributions were successfully normalized with logarithmic transformations, and the inter-version and test-retest correlations were recalculated. As these results were practically almost identical (i.e., within ±0.01 units) with those obtained with the original measure, only the results with the original measures are reported.

3 Results

3.1 Task performance in FSD/HMD and learning effects

Table 1 shows the effects of version (FSD/HMD) and time (first/second) on EPELI task performance measures and related descriptive statistics. Children achieved higher Total scores, TBPM scores, and Task efficacies in the FSD version with small effect sizes. They made almost twice as many clock checks in the FSD version compared to the HMD version, which is in line with their better TBPM performance in the FSD version. To inspect this phenomenon further, we reran the analysis by using clock-viewing duration (i.e., the total duration of clock-viewing in seconds) as the dependent variable and found a medium-sized version effect (t (69.155) = 4.544, $p < .001$, $d = 0.55$). As Total score also includes the TBPM tasks, we did a *post hoc* analysis for Total score without the TBPM tasks. This analysis found both effects of version (t (67.40) = 2.642, $p < 0.01$, $d = 0.32$) and time (t (67.38) = 6.786, $p < 0.001$, $d = 0.83$), which suggests that the difference in Total score between the versions is driven not only by a better TBPM performance.

In the second session, the children achieved higher Total scores (large effect size), higher TBPM scores (medium effect size), and higher EBPM scores (small effect size). They also performed more actions and navigated more efficiently, for which the effect sizes were small. However, Task efficacy did not change, indicating that they also did more irrelevant actions during the second session compared to the first. This is reflected in the fact that the number of irrelevant actions (i.e., actions that do not work towards given goals) as analyzed separately also increased from the first session to the second (t (70) = 3.501, $p < 0.001$, $d = 0.40$). Because learning effects were found in five variables, we checked with *post hoc* analyses if their magnitude was different depending on which version was performed first. The learning effect was larger after the HMD version than after the FSD version for Total score (mean change: after HMD 8.42, after FSD 3.18; t (63.72) = 3.477, $p < 0.001$, $d = 0.44$) and TBPM (mean change: after HMD 3.55, after FSD 1.00; t (64.50) = 3.395, $p = 0.001$, $d = 0.42$). For other measures, the learning effect was not affected by the version used in the first measurement.

3.2 Subjective experiences in FSD/HMD

The results of the Presence questionnaire, which was used to examine differences in subjective experience between the EPELI

10.3389/frvir.2023.1138240

TABLE 1 The effects of version (FSD/HMD) and time (1st/2nd session) on EPELI task performance measures and related descriptive statistics.

Dependent variable	Descriptive statistics				Mixed model test statistics							
	Version, mean (SD)		Session, mean (SD)		HMD vs. FSD				1st vs. 2nd session			
	HMD	FSD	1st	2nd	Estimate (SD)	*t*	*p*	*d*	Estimate (SD)	*t*	*p*	*d*
Total score (0–70)	54.42 (6.77)	57.13 (7.02)	52.94 (7.17)	58.75 (5.46)	**2.685 (0.749)**	**3.588**	**<0.001**	**0.44**	**5.814 (0.749)**	**7.765**	**<0.001**	**0.95**
Task efficacy	0.43 (0.13)	0.47 (0.10)	0.46 (0.12)	0.44 (0.12)	**0.031 (0.015)**	**2.103**	**0.039**	**0.26**	-0.019 (0.015)	-1.272	0.208	-0.16
Navigation efficacy	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)	-0.001 (0.002)	-0.444	0.658	-0.05	**0.007 (0.002)**	**2.727**	**0.008**	**0.33**
Controller motion	-	-	-	-	-	-	-	-	-	-	-	-
Total actions	456.91 (151.9)	426.1 (109.63)	422.11 (124.90)	461.53 (138.50)	-26.492 (16.958)	-1.562	0.123	-0.20	**43.256 (16.961)**	**2.550**	**0.013**	**0.32**
TBPM (0–13)	6.35 (3.03)	7.67 (2.69)	5.92 (2.91)	8.16 (2.49)	**1.317 (0.371)**	**3.554**	**<0.001**	**0.43**	**2.261 (0.371)**	**6.095**	**<0.001**	**0.74**
Clock checks	34.65 (11.77)	65.19 (33.84)	48.82 (31.58)	51.29 (27.53)	**30.289 (3.933)**	**7.701**	**<0.001**	**0.95**	2.918 (3.934)	0.742	0.461	0.09
EBPM (0–6)	4.13 (0.75)	4.29 (0.7)	4.01 (0.77)	4.41 (0.63)	0.157 (0.1)	1.574	0.120	0.19	**0.396 (0.100)**	**3.964**	**<0.001**	**0.48**

N = 72. The effects that are significant at the level of $p < .05$ are written in bold. Cohen's *d*, effect size.

TABLE 2 The linear mixed models with the Presence questionnaire items as dependent variables and EPELI version (HMD/FS) and time (1st/2nd session) as fixed factors.

	Descriptive statistics					Mixed model test statistics						
	Version, mean (SD)		Session, mean (SD)			HMD vs. FSD				1st vs. 2nd session		
Question	HMD	FS	1st	2nd	Estimate (SD)	*t*	*p*	*d*	Estimate (SD)	*t*	*p*	*d*
1. How natural did your interactions with the environment seem?	4.61 (1.19)	4.41 (1.66)	4.66 (1.40)	4.37 (1.47)	-0.210 (0.200)	-1.049	0.298	-0.13	-0.300 (0.200)	-1.503	0.138	-0.18
2. How much did the environment involve you?	5.22 (1.3)	4.76 (1.73)	5.13 (1.61)	4.87 (1.45)	**-0.430 (0.211)**	**-2.035**	**0.046**	**-0.25**	-0.251 (0.211)	-1.19	0.238	-0.14
3. How natural was the mechanism which controlled movement through the environment?	3.93 (1.76)	3.74 (1.76)	3.84 (1.74)	3.84 (1.78)	-0.170 (0.215)	-0.788	0.434	-0.10	-0.009 (0.215)	-0.041	0.968	-0.01
4. How much did your experiences in the virtual environment seem consistent with your real-world experiences?	5.12 (1.39)	4.39 (1.67)	4.81 (1.49)	4.74 (1.66)	**-0.770 (0.195)**	**-3.942**	**<0.001**	**-0.50**	-0.074 (0.195)	-0.378	0.707	-0.05
5. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?	2.22 (1.28)	1.70 (1.01)	2.03 (1.22)	1.91 (1.15)	**-0.547 (0.168)**	**-3.258**	**0.002**	**-0.39**	-0.098 (0.168)	-0.582	0.563	-0.07
6. How much did the control devices interfere with the performance of assigned tasks or with other activities?	1.58 (1.10)	1.55 (1.18)	1.59 (1.19)	1.54 (1.09)	-0.037 (0.194)	-0.191	0.849	-0.02	-0.045 (0.194)	-0.231	0.818	-0.02
7. How well could you concentrate on the assigned tasks or required activities?	5.76 (1.05)	5.30 (1.42)	5.51 (1.24)	5.57 (1.29)	**-0.465 (0.183)**	**-2.534**	**0.014**	**-0.30**	0.078 (0.183)	0.427	0.671	0.05
8. How well could you hear sounds?	6.79 (0.60)	6.76 (0.66)	6.71 (0.73)	6.84 (0.50)	-0.036 (0.107)	-0.338	0.736	-0.03	0.138 (0.107)	1.285	0.201	0.11
9. Were there moments during the virtual environment experience when you felt completely focused on the task or environment?	4.74 (1.94)	4.36 (2.07)	4.5 (2.00)	4.61 (2.02)	-0.364 (0.306)	-1.189	0.239	-0.14	0.131 (0.306)	0.43	0.669	0.05
Three additional questions that were not in the original Presence Questionnaire 3.0:												
10. How enthusiastic did you feel about the tasks?	5.58 (1.26)	5.23 (1.57)	5.56 (1.40)	5.27 (1.44)	-0.297 (0.149)	-1.994	0.050	-0.24	-0.290 (0.149)	-1.95	0.055	-0.24
11. How interesting did the tasks seem to you?	5.14 (1.39)	4.56 (1.74)	5.13 (1.62)	4.60 (1.52)	**-0.533 (0.211)**	**-2.522**	**0.014**	**-0.31**	**-0.526 (0.211)**	**-2.491**	**0.015**	**-0.30**
12. How much effort did you put into your performance?	6.25 (0.88)	6.15 (0.93)	6.18 (0.96)	6.23 (0.85)	-0.101 (0.134)	-0.758	0.451	-0.09	0.052 (0.134)	0.392	0.697	0.05

$N = 72$. All questions were answered from a Likert scale with a range of 1–7. The effects that are significant at the level of $p < .05$ are written in bold. Cohen's *d*, effect size.

10.3389/frvir.2023.1138240

versions, are displayed in Table 2. The children evaluated that in the HMD version, the environment involved them more (small effect size), their experiences felt more consistent with the real world (medium effect size), they could concentrate better on the assigned tasks (small effect size), and the task seemed more interesting compared to the FSD version (small effect size). They also reported more problems in the display quality after the HMD than the FSD version (small effect size), but the problems were minor in both versions (HMD mean 2.22 and FSD mean 1.70 on a scale of 1–7). There were no differences between the two sessions, except that the children evaluated the tasks as appearing more interesting after the first session than the second (small effect size). The children reported very few potential cybersickness symptoms after both the HMD (mean sum 0.83 on a 0–14 scale) and FSD version (mean sum 0.56), and there was no difference between the versions ($V = 358.5$, $p = 0.07$). When asked to compare the two versions after the second EPELI session, most children evaluated the HMD version as being more realistic (48 out of 51, exact binomial test, $p < 0.001$) and preferable (36 out of 48, exact binomial test, $p < 0.001$) than the FSD version. Majority of the children (31 out of 49) also evaluated the HMD version as being the easier to play, but this difference was not significant (exact binomial test, $p = 0.09$).

3.3 Similarities and differences between experimented-supervised laboratory testing and parent-supervised home testing

The groups who performed FSD-EPELI either supervised by experimenter in laboratory or by parent at home displayed very similar results, as there were no group differences in task performance (Supplementary Table S3) or perceived presence (Supplementary Table S4). There were no differences regarding age, handedness, gender, parental education, or family income between the laboratory testing and home testing groups either (Supplementary Table S1).

3.4 Associations between EPELI efficacy measures and BRIEF

The correlations between EPELI efficacy measures and BRIEF across EPELI versions (FSD/HMD) and sessions (first/second) are shown in Table 3. BRIEF GEC correlates with both Task efficacy ($r = -0.37$) and Navigation efficacy ($r = -0.33$) on the first session, but not on the second. To interpret this result, we computed correlation between BRIEF GEC in the two sessions and found that association to be strong ($r = 0.77$, $t\ (67) = 9.905$, $p < 0.001$). To evaluate how carefully the parents had considered their answers on each test session, we also compared the testing times and found out that parents had used less time on the second test session (median time between opening the questionnaire and closing it, with 9.00 min for the 1st session, and 7.13 min for the 2nd session, $U = 3166$, $p = 0.014$). When the correlations are inspected with each version at a time but including both assessment sessions, only Navigation efficacy is associated with BRIEF.

3.5 Inter-version correlations and test-retest stability

Inter-version and test-retest correlations for the eight EPELI measures are presented in Table 4, and distributions of the EPELI variables in both versions are shown in Figure 3. Regarding partial correlations across EPELI versions, the highest were found in Total score, Task efficacy, and Total actions (0.43–0.52), followed by Navigation efficacy, Controller motion, TBPM, EBPM, and Clock checks (0.29–0.40). The highest partial correlations across test sessions were obtained in Total score, Task efficacy, and Total actions (0.43–0.54), followed by Navigation efficacy, TBPM, EBPM, Controller motion (0.31–0.39). Clock checks was not correlated across test sessions. As the effects of version (HMD/FSD) and session (first/second) were also analysed for clock-viewing duration (see 3.1), we also calculated the inter-version and test-retest correlations for this measure and found it to be correlated both between test versions (partial $r = 0.46$, $p < 0.001$) and test sessions (partial $r = 0.28$, $p < 0.05$).

4 Discussion

Rapid advances in VR display technology now allow researchers and clinicians to choose from a wider set of technical platforms than before. This has created a need to inspect the strengths and weaknesses of each platform and to compare the results they yield. To this end, the current study set out to compare the HMD and FSD versions of EPELI, a naturalistic task of goal-directed behavior.

Overall, the results attest to the viability of both hardware implementations, as task performance and subjective experience ratings were to a large extent comparable, and all task performance measures were correlated across the two versions. We also found some differences between the versions but with mostly small effect sizes. Most notably, children's performance was somewhat better on FSD, while the HMD version was preferred and received better evaluations on several questions related to user experience. There were no differences between the parent-supervised home group and the examiner-supervised laboratory group in FSD-EPELI, which supports its feasibility for remote testing. Both versions are associated with parent-evaluated problems of executive function on the first assessment, but interestingly, not on the second one that took place several months later. All in all, both versions have their own benefits, such as the more sophisticated body movement tracking and higher immersiveness in the HMD version, and opportunities to improve cost-effectiveness and reachability via home-based testing with the FSD version. Below, we discuss each key finding in greater detail.

4.1 Task performance in the FSD and HMD versions of EPELI

Although the level of task performance was similar in the two versions, some modest but noticeable differences emerged. The children achieved higher Total and TBPM scores and task efficacies in the FSD version but with small effect sizes. This is

TABLE 3 Correlations of EPELI efficacy measures and BRIEF.

EPELI measure, session, version	BRIEF GEC
	r
Task efficacy, 1st session, both versions	-0.37**
Navigation efficacy, 1st session, both versions	-0.33*
Task efficacy, 2nd session, both versions	0.03
Navigation efficacy, 2nd session, both versions	-0.15
Task efficacy, both sessions, HMD version	-0.18
Navigation efficacy, both sessions, HMD version	-0.37**
Task efficacy, both sessions, FSD version	-0.18
Navigation efficacy, both sessions, FSD version	-0.11

N = 69–72. ***p < .001, **p < .01, *p < .05, based on *p*-values with False Detection Rate correction.

TABLE 4 EPELI measure intercorrelations between the HMD- and FSD-versions and the first and second sessions.

Measure	HMD vs. FSD-version			1st vs. 2nd session		
	r	ICC[a]	*partial* r[b]	*r*	ICC[a]	*partial* r[d]
Total score	0.25	0.23*	0.52***	0.47***	0.32***	0.54***
Task efficacy	0.47***	0.46***	0.48***	0.45***	0.46***	0.47***
Navigation efficacy	0.36**	sing.[c]	0.40**	0.39**	0.38***	0.39**
Controller motion	0.33**	0.15**	0.34**	-0.16	sing.[c]	0.31*
Total actions	0.37**	0.38**	0.43***	0.42***	0.41***	0.43***
TBPM	0.12	0.11	0.32*	0.25	0.19*	0.32**
Clock checks	0.27*	0.10	0.29*	0.15	sing.[c]	0.18
EBPM	0.21	0.21	0.30*	0.29*	0.25**	0.31*

N = 67. ***p < .001, **p < .01, *p < .05, based on *p*-values with False Detection Rate correction.
[a]= ICC2,1.
[b]= partial correlations with time (1st or 2nd session) as a covariate.
[c]= singularity error.
[d]= partial correlations with version (HMD, or FSD) as a covariate. TBPM, time-based prospective memory score; EBPM, event-based prospective memory score.

in line with some previous research suggesting that even though HMDs might produce a superior feeling of presence, the use of FSDs can in some cases lead to better performance outcomes (Makransky et al., 2019; Barrett et al., 2022). This raises interesting questions regarding the role of immersiveness in the measurement of cognitive performance. Considering the case with EPELI, it is important to note that the task instructions are given orally, and the audio is delivered in a similar way between the versions (i.e., using stereo headphones, adequate volume, and by placing each sound source in the stereo image in the place that corresponds to its location in virtual space). The children reported that they experienced no problems in hearing the sounds in either version. Even though the dragon character can be seen talking, the facial expressions and mouth movements are not synchronized with the words, which means that looking at the dragon does not necessarily help in memorizing the instructions. Nevertheless, the children can look around (but not walk around) in the VR environment while listening to the instructions and might be more tempted to do so with HMD, as they report the HMD version to be more involving. This could mean that in the HMD version, they focus less on listening to the instructions and more on irrelevant but appealing visual stimuli in the environment, which would lead to better performance with FSD. During the execution phase that follows the instructions, as well as the instruction phase itself, the more immersive experience of the HMD version could lead them to be drawn more strongly to the irrelevant stimuli. Supported by their eye tracking data, Barrett and others (2022) speculate that just looking around might be more fun with HMD as compared to FSD, which would be consistent with this explanation. As there was no difference in reported effort between the versions, and as the children reported the tasks as more interesting when using HMD, lack of motivation during HMD performance is unlikely to explain the better performance in FSD. Therefore, a logical

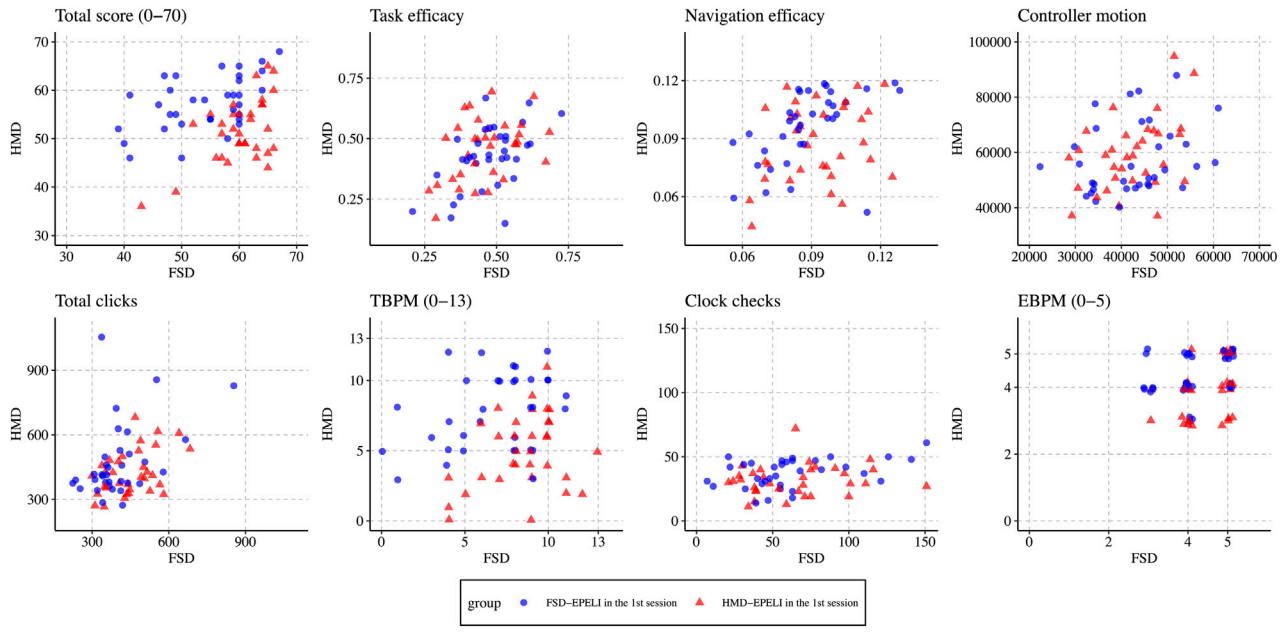

FIGURE 3
Distributions of EPELI variables in HMD and FSD. TBPM, Time-based prospective memory score. EBPM, Event-based prospective memory score.

explanation for these inter-version differences could be that the higher immersiveness of the HMD version more easily disengages the participants from listening to instructions for the given tasks and performing them, which leads to worse performance.

Another potential reason for the differences in Total score, TBPM score, and Task efficacy between the versions lies in the differences in the control interfaces (i.e., using head movements and a hand controller vs. traditional devices, mouse/trackpad). However, several pieces of evidence render this explanation unlikely. First, the children reported very few problems with the control devices, with no differences between the two versions. Second, the children familiarized themselves with the controls during the demo section of EPELI, and it was ensured that all participants could perform the required actions as needed. Third, EPELI does not place heavy time constraints on the participant, as there is sufficient time to perform all required actions even at a relaxed pace, if one keeps focused on them and avoids getting into task-irrelevant behaviors that the naturalistic environment allows. This means that no quick actions or particularly skillful use of the control devices are needed to perform well in EPELI. Fourth, although inter-version differences favored FSD-EPELI in these three measures, most of the children evaluated that HMD-EPELI was easier to play. Thus, it is unlikely that the differences in the control interfaces would have played a major role in the results.

The previous studies comparing mental load between FSD and HMD conditions provide ground for speculating upon possible cognitive processes behind the inter-version differences. At least two studies that compared FSD and HMD conditions found higher mental load when using an HMD, either based on self-report (Brooks et al., 2017) or EEG responses (Makransky et al., 2019). In contrast with these findings, other studies have found no differences in self-reported mental effort and psychophysiological responses (Chang et al., 2020) or total EEG activation (Li et al., 2020). It should be noted that Li and others (2020) used quite narrow FOV (<20°) that was the same for both conditions, which might explain the lack of differences. In our study, a possible higher cognitive load with the HMD could have been induced by extraneous visual information due to its larger FOV and stereoscopic view. In EPELI, most visual stimuli in the environment are irrelevant to the tasks at hand and have the potential to distract the participant from performing these tasks. Therefore, in the HMD version the load on the bottom-up visual processes could be higher and thus might cause more interference with the top-down cognitive processes (e.g., working memory) required to perform the instructed tasks (see, e.g., Repovš & Baddeley, 2006). This line of thought is compatible with the interpretations given above.

The differences between the two hardware versions were particularly prominent in time monitoring. The number of clock checks in FSD-EPELI was almost double that in HMD-EPELI, which corresponds to a large effect size. At least three explanations could account for this finding. First, if the suggestion above regarding the higher bottom-up visual processing load with HMD is correct, less cognitive resources could be available for time monitoring in the HMD version. This explanation is compatible with previous research showing that increasing the cognitive demands of the ongoing task in a prospective memory paradigm can result in less active time monitoring (Khan et al., 2008). Second, it takes less effort to check the time in FSD-EPELI, as the watch can be viewed in the lower right corner of the screen with a single click. In HMD-EPELI, the participant needs to raise or turn his/her arm slightly and look towards the hand controller in virtual space to see the

watch, like checking the time from a wristwatch. This might reduce the tendency to check the time. Third, while in the HMD-EPELI the hand controller does not display the watch or a white circle until purposefully raised and then kept still for a second, in FSD-EPELI a white circle where the watch appears is displayed in the lower right corner of the screen also when the clock is not shown, which might serve as a cue for time monitoring (see Figure 1D). These three explanations do not rule out each other and all might play some role behind the differential time monitoring in the two versions. Future research should establish which factors contribute most to the observed differences between the HMD and FSD versions.

One should note here that the inter-version effect of Total score remained even when TBPM score was subtracted from the Total score. Thus, the inter-version differences in Total score and Task efficacy are unlikely to be driven only by the more accurate time monitoring performance in FSD.

4.2 Learning effects between 1st and 2nd sessions

There was a large learning effect on the Total score and TBPM score and a small one on the EBPM score from the first session to the second, even though the interval between the sessions was long, over 7 months on average. Given that verbal word list or story learning tests are often reported to yield at least moderate learning effects when the same material is used on both sessions (e.g., Wechsler, 1997; Woods et al., 2006), it is not surprising to find some learning effects here. The learning effect might have been amplified by the fact that EPELI involves tasks that are acted upon rather than just orally repeated, as practice effects in neuropsychological tasks have been hypothesized to be related not only on declarative (e.g., remembering the test items) but also to procedural (e.g., remembering how to perform the test) memory (Duff, 2012). The learning effects observed in the present study did not cause ceiling effects on the second assessment and therefore do not compromise EPELI's utility for test-retest settings. With a commonly used word list learning task, The California Verbal Learning Test, the learning effects are notably smaller when alternative materials are used on the second assessment (Woods et al., 2006). This suggests that with EPELI, alternative scenarios with different task instructions should be used in retest situations where a minimal learning effect is desired.

The children also performed more actions during the second assessment than in the first. This small effect is partly explained by a better Total score, but the amount of irrelevant behavior also increased between the sessions. It could be that when performing EPELI the second time, the children were more prone to experiment with the environment freely and less compelled to limit themselves only to the instructed tasks. However, this possible change from the first session to the second one was not reflected in their self-reported effort, which stayed constant between the sessions.

The children navigated more efficiently in the second session, which could reflect that they had become more familiar with the apartment through practice and were therefore able to plan their routes more efficiently. Even so, there was no difference in Task efficacy (i.e., the efficacy in interacting with the objects) between the sessions. This is because the children not only performed more of the given tasks in the second session than in the first, they also engaged more in extraneous, task-irrelevant behavior. However, based on more efficient navigation, this extraneous behavior did not include excessive walking around the apartment, but was present only in the interactions with the objects.

4.3 Subjective experiences

We found that when compared with FSD, the HMD environment was perceived as more involving, the experiences felt more consistent with the real world, the children reported that they were able to concentrate better on the given tasks, and the tasks seemed more interesting to them. When asked to compare the two versions directly, the children evaluated HMD-EPELI as being more realistic and preferable. These findings are consistent with studies on commercial games finding that HMD elicits a stronger sense of presence (see Caroux, 2023, for a meta-analysis) and immersion, and a greater arousal of positive emotions (e.g., Tan et al., 2015; Pallavicini et al., 2018; Pallavicini et al., 2019; Pallavicini & Pepe, 2019), as well as user satisfaction (Shelstad et al., 2017) than FSD-based hardware. Also Makransky and others (2019) reported that students felt more present in HMD than in FSD condition during a learning task. Using driving simulation with an embedded Stroop task, Chang and others (2020) found that students reported an HMD to be easier to use than an FSD. This was echoed in our data as most of the children evaluated HMD-EPELI as being easier to play than FSD-EPELI, although this difference was not statistically significant. In case this is a true effect, it could relate to interaction with the environment being more naturalistic in the HMD version than in the FSD version, which could be achieved with the position-tracked hand controller. For example, checking the time in the HMD version took place by looking at the controller similar to looking at a wristwatch and playing the drums by swinging the controller at them like using a drumstick, as opposed to performing these actions merely by clicking the mouse button.

Regarding the two sessions, the only difference in subjective experiences was that the children reported the tasks as being more interesting in the first session than on the second one. Still, their average evaluation as to how interesting the tasks felt remained high in the second session.

The children reported very few sickness symptoms after either version. This is an important finding given the negative association between cybersickness and the sense of presence (Weech et al., 2019) and in line with the results in our earlier results, some of them obtained with a different HMD model, Pico Neo 2 Eye (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). Overall, the current HMD systems seem to be able to offer an enjoyable VR experience without cybersickness symptoms (Kourtesis et al., 2019). This is of course not to say that sickness symptoms would always be absent under all conditions. Using a target detection task involving flying, Brooks and others (2017) reported higher mental workload and discomfort in HMD compared to FSD. Given the results described above, their findings might stem from the fact that somewhat older and less sophisticated HMD hardware (NVISOR ST50) was used. Also, flying simulations might be more prone to cause sickness symptoms as compared to EPELI where the movements are self-paced and walking is done via

teleporting. Therefore, when developing new VR tasks, researchers should continue to evaluate any possible sickness symptoms carefully and, if needed, modify their tasks to eradicate these symptoms.

The current study employed Oculus Go HMD hardware released in 2018, and more technically advanced models have since been introduced to the market. In a previous study, we employed both Oculus Go and a more advanced Pico Neo 2 Eye HMD and found no differences on the same Presence questionnaire that was used in this study, except for fewer problems with the hand controller for the Pico (Seesjärvi et al., 2022a). As these problems were on average very few for both models, the findings between the different HMD hardware we have used can be considered very similar. Taken as a whole, we expect the perceived presence to remain quite the same if similar HMD equipment with slightly different specifications is used. This being said, more realistic interaction methods, such as hand tracking-based object manipulation, walking based on a treadmill instead of teleporting, and augmented reality setups could lead to an enhanced feeling of "being there". Further research on perceived presence is therefore warranted when such advancements are adopted.

Previous studies with different tasks provide useful insights on why different implementations of the same task could prove useful in different situations. The Multiple Errands Test was first developed to be performed in a real-life shopping precinct (Shallice & Burgess, 1991) and later modified for different settings, such as a hospital (see the review by Rotenberg et al., 2020). Later, several desktop FSD variations that have the benefit of greater experimental control but could be less ecologically relevant as being less presentative and more removed from the everyday environments, have been developed (e.g., Rand et al., 2009; Jovanoski et al., 2012; Raspelli et al., 2012; Cipresso et al., 2014). Recently, Webb et al. (2021) created a simplified tablet version, OxMET, to be used as a brief screening tool. Even though this tablet version differs more from the original real-world MET than the desktop versions (e.g., it does not include any walking in a three-dimensional environment, but the participant navigates by touching pictures of shops in a cartoon shopping street), it has strong potential for its intended use, that is, as a screening tool for the executive problems that this kind of naturalistic paradigm aims to capture. For EPELI, the FSD version allows, for example, large-scale data collection, even though it might offer less natural sensorimotor contingencies and therefore be less immersive than the HMD version. The development of the VR-EAL, which is a neuropsychological test battery implemented by using immersive HMD-VR (Kourtesis et al., 2021), provides another interesting comparison point on this theme. The VR-EAL, which has been developed for adults, has some qualities that might enable it to provide better sensorimotor contingencies than the current EPELI HMD version, as it uses a combination of physical movement and teleportation as a navigation method, and it is performed in an upright position instead of a sitting position. For safety reasons (see Seesjärvi et al., 2022a), we have chosen to use only teleportation and a sitting position with school-aged children. The interface system of the VR-EAL reflects some of the more advanced possibilities of immersive VR systems, and a possible FSD version of the same paradigm using a keyboard or mouse as the interaction method might be markedly less immersive. However, such a version could surely be implemented, and would probably provide new use cases for the VR-EAL as well.

4.4 Comparisons between laboratory and home testing

The fact that there were no differences between laboratory and home testing in task performance or subjective presence ratings supports the feasibility of parent-supervised remote testing. In line with this, Zuber et al. (2021) found that the laboratory and online versions of a prospective memory task, Geneva Space Cruiser, yielded similar results in an adult lifespan sample. The findings of Backx et al. (2020) with several tests from Cambridge Neuropsychological Test Automated Battery are also promising for remote testing, as no differences were found in the performance indices between laboratory and home testing. It should be noted that Backx et al. (2020) found the reaction times to be slower at home than in the laboratory, which they suggested was caused by variation in the computer hardware. This should be kept in mind when new measures are developed for EPELI, but as the present EPELI performance variables did not contain any reaction time measures or other indices that would be very sensitive to subtle variations in time measurement, it is not a concern here.

Thus, the present study shows that laboratory and remote testing can produce comparable results also in naturalistic tasks such as EPELI. It is also important to note that the present participants were children who could be more prone than adults to perform differently when not supervised by an experimenter. As the COVID-19 pandemic has shown, unexpected events with tremendous impacts on societies are possible and can challenge the routines of scientific research and clinical work. Hence, the present findings on the feasibility of remote testing are very timely and have broader relevance for cognitive assessment. As remote assessments save time and resources both for the assessor and the assessee, they are very likely to become even more common in future.

4.5 Associations between EPELI efficacy measures and BRIEF

The current study indicates that when administered the first time, FSD-EPELI efficacy measures are also associated with parent-rated problems of executive function (BRIEF), as previously shown for HMD-EPELI (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b). However, we were surprised to find that these associations disappeared in the second assessment. As BRIEF was strongly correlated between the sessions and parents seem to have used, on average, adequate time to fill out the questionnaire on both occasions, this change was likely to be caused mostly by a change in children's behavior in EPELI from the first session to the second. As children engage in more actions on the second session, one possible explanation could be that in the second session, also those children who do not exhibit executive problems in everyday life resort to extraneous behavior more easily, which makes Task efficacy less representative of these problems. Another possibility is that it is the novelty of the task that makes Task efficacy representative of executive function problems in the first session. This relates to the finding that the involvement of executive functions is considered to be highest when the task is new (Rabbit, 2004). Barrett et al.

(2022) speculate that with more exposure to a VR interface beyond the first session, the novelty and thereby the initial enthusiasm would diminish and thereby alter the results in following sessions. Whatever the explanation turns out to be, these results should be kept in mind when using EPELI in test-retest settings or in longitudinal studies.

4.6 Inter-version correlations and test-retest stability

Earlier, we showed that HMD-EPELI has acceptable internal consistency in six out of eight measures (Seesjärvi et al., 2022a). The internal consistency was highest (Cronbach's α = 0.83–0.88) for the measures with the most data points, i.e., Controller motion, Total actions, and Task efficacy which is closely related to Total actions. The internal consistency was acceptable for Total score, Navigation efficacy, and Clock checks (α = 0.70–0.74). Here, all these six measures were associated between HMD- and FSD-EPELI (partial r = 0.29–0.52) and all except the number of clock checks were correlated between sessions (partial r = 0.31–0.54), which attests to their stability across the two versions and a time interval that on average spanned over 7 months.

As a comparison point, Backx et al. (2020) reported correlations of ρ = 0.39–0.73 between laboratory and home assessment sessions 1 week apart with several indices from the Cambridge Neuropsychological Test Automated Battery. Using a prospective memory task with a typical dual-task paradigm, Zuber et al. (2021) found correlations of r = 0.56–0.68 between laboratory and home assessment sessions 1 week apart for three indices (ongoing task score, prospective memory performance, time monitoring), and correlations of r = 0.66–0.78 for another sample where two sessions were done in the laboratory 1 week apart. Using the somewhat more complex prospective memory task of Virtual Week, Mioni et al. (2015) found varied correlations (r = 0.13–0.74) with a time interval of 1 month. These earlier studies indicate that the test-retest correlations can vary considerably based on, among other things, task complexity, type of the measure, and environment (laboratory/home).

When considering test-retest stability, it should be noted that although the two EPELI versions are highly similar, they are not identical which is the case in some other studies referred here (Backx et al., 2020; Zuber et al., 2021). Also, the average time interval between the sessions was exceptionally long, spanning over 7 months. The test-retest stability has been found to decrease with increasing test-retest intervals (Duff, 2012), as the correlation might not reflect only measurement error but also true change. It should also be noted that the complex nature of executive function tasks, which involve multiple cognitive processes, could make them more prone to performance variability (Delis et al., 2004). To acquire an estimate of test-retest stability of each EPELI version that would be more comparable to the earlier literature, future studies should be conducted using only one version at a time, doing both sessions in the laboratory, and with markedly shorter time interval between the sessions. Even though the long interval between the sessions makes comparison with the earlier literature more difficult, the current study has the benefit that it proves that such test-retest correlations do exist even after such a long delay.

Time monitoring, when measured as the number of clock checks, was correlated between the versions, but not between the sessions. It is worth noting here that the clock checking mechanism is different between the versions (raising the controller and looking at it in the HMD condition, pressing a button to view the time in the lower right corner of the screen in the FSD condition) and only in the FSD version a white circle remains in the lower right corner of the screen where the clock will appear, which can work as a cue for time checking. Also, it has been found that time perception can be compressed with an HMD as compared to FSD condition (Mullen & Davidenko, 2021). If this phenomenon varies from individual to individual, it can weaken the correlation of time monitoring measures in the two conditions. Children might also engage in different time monitoring strategies between immersive VR environments and conventional FSD conditions. Further test-retest studies using only one of the two versions will reveal to what extent the lack of test-retest correlation in the number of clock checks here is tied to the differences between HMD and FSD conditions. Lastly, it should also be noted that the duration of clock-viewing correlated both between the versions and between the sessions. Therefore, in a naturalistic prospective memory task like EPELI, using clock-viewing duration instead of the number of clock checks might be a more robust way to measure time monitoring behavior.

Regarding prospective memory, both TBPM and EBPM scores were correlated between the versions and sessions (partial r = 0.30–0.32). Thus, even though the internal consistency of these measures was earlier found to be poor (Seesjärvi et al., 2022a), it appears that these measures are relatively stable across time. Using Virtual Week, which is another attempt at a more ecologically valid test for prospective memory, Mioni et al. (2015) reported lower or similar test-retest correlations for EBPM (r = 0.13–0.40) but higher for TBPM (r = 0.58–0.74). Further attempts should be made to improve the psychometric properties of naturalistic prospective memory tasks, as prospective memory is important for self-dependent everyday functioning.

4.7 The future potential of EPELI for healthcare settings

Considering the potential future clinical use of EPELI, some discussion regarding the steps already taken in this direction is in order. In a joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology, Bauer et al. (2012) identify eight key issues relevant to the development and use of computerized neuropsychological assessment devices in healthcare settings. These eight issues concert marketing and performance claims; end-user requirements; hardware/software/firmware issues; privacy/data security/identity verification/testing environment; reliability and validity; cultural/experiential/disability factors; use of computerized testing and reporting services; and the need to control for response validity and effort. Regarding reliability, we have previously shown with the HMD version that most EPELI measures show acceptable internal consistency (Seesjärvi et al., 2022a). As regards to validity, EPELI shows predictive and discriminant validity in differentiating between

children with ADHD and typically developing controls (Merzon et al., 2022; Seesjärvi et al., 2022b) and ecological validity (veridicality) by correlating with parent-rated problems of everyday executive function both in children with ADHD (Seesjärvi et al., 2022b) and typically developing children (Seesjärvi et al., 2022a). The children reported only negligible cybersickness symptoms in this and previous studies (Merzon et al., 2022; Seesjärvi et al., 2022a; Seesjärvi et al., 2022b), and all were able to learn the controls and perform the whole task. The current study shows that the FSD version has potential to be used remotely, as the performance and subjective ratings of home and laboratory groups were equal. However, several issues remain to be addressed. The marketing claims of any potential product should be based on solid scientific findings. The end-user (i.e., the assessor) requirements should be defined while considering the required knowledge of psychological assessments and technical competence, and the results should be represented in a clear format that is easy to interpret. The possible online data storage needs to be implemented by using proven and secure platforms. Regarding this, a fully online study that used an adult version of EPELI and Microsoft Azure platform has already been successfully performed (Jylkkä et al., 2023). In case EPELI versions for other languages and groups of other age groups (e.g., older adolescent) and clinical diagnoses (e.g., brain injury or dementia) are developed, their feasibility should be examined separately. Kourtesis and MacPherson (2021) have pointed out with their work with VR-EAL and young adults that immersive VR paradigms have the potential to meet all the key criteria mentioned by Bauer et al. (2012). A similar work that considers all the eight issues simultaneously should be pursued for any potential healthcare version of EPELI. To our knowledge, such a work would be the first to consider all these aspects in a task that can be used to study goal-directed behavior of children in immersive VR. Kourtesis and MacPherson (2021) suggest that a future version of VR-EAL should consider hand and head movement to evaluate whether the examinee is motivated to engage with the given tasks. This should be attempted with EPELI as well, as the effort level has been found to substantially affect performance on neuropsychological tests (Constantinou et al., 2005; Stevens et al., 2008; West et al., 2011).

4.8 Limitations and future directions

As with all research, this study has its limitations. The delay between the two assessment sessions was unusually long, on average over 7 months, caused by restrictions imposed by the COVID-19 pandemic. This hinders the comparison with earlier literature, as the present inter-version and test-retest correlations could be weaker and learning effects more modest than in studies with considerably shorter time intervals. With a shorter delay, the inter-version associations could be evaluated more accurately. On the other hand, the delay employed here comes closer to typical minimum interval between clinical neuropsychological assessments, which is usually at least a year in children, even though we are not aware of empirical data that would allow the development of guidelines for generally acceptable minimum test-retest intervals in clinical settings (see Heilbronner et al., 2010). Therefore, different time intervals come with different strengths and drawbacks. In the current study, the primary aim was to study the associations between the two versions using a within-subject design and at the same time acquire some estimates about test-retest stability. To acquire true test-retest correlations of each EPELI version, further studies should be conducted with a single version repeatedly taken by each participant. As for many other neuropsychological tests tapping memory, it could be that an optimal version of EPELI for multiple measurements within the same individual would require parallel versions, that is, several task sets of equal difficulty.

In future, the FSD version should also be employed to study clinical groups, such as ADHD and autism spectrum disorders. Our previous research has shown robust and distinctive differences between children with ADHD and matched controls in HMD-EPELI (Merzon et al., 2022; Seesjärvi et al., 2022b). As the current study attests to the feasibility of children's FSD-EPELI for parent-supervised remote testing, an online study with a markedly larger dataset could be pursued. In this first study with the FSD version, all families that performed it remotely at home could do so with the given instructions and no differences between the groups that performed FSD-EPELI either at home (supervised by a parent) or at lab (supervised by a researcher) were observed. The children also reported high ratings on questions regarding their enthusiasm, how interesting EPELI was and how much effort they put into their performance. Also the ratings considering the display/control device quality were favorable for both versions. However, the usability and acceptability of different EPELI versions were not fully probed, which should be conducted in future studies.

New technologies continue to emerge rapidly, which calls for continuing the research on human-computer interfaces. This study was limited to two technical configurations, while there would be many possible alternatives to be tested. Thus far, we have chosen to ask the children to perform HMD-EPELI in a sitting position, as during the piloting stages of our earlier studies (Seesjärvi et al., 2022a; Seesjärvi et al., 2022b) especially younger children with no prior VR experience had problems playing in a standing position (e.g., they tried to reach something to lean on) and some reported feeling slightly dizzy (see Seesjärvi et al., 2022a). This decision made to ensure participant safety, renders performance on the HMD and FSD versions more similar, but a standing position could improve sensorimotor contingency in HMD-EPELI. As even more natural VR technologies and human-computer interfaces emerge, their benefits for naturalistic cognitive tasks should be examined, too. As an example, these technologies include using hand position tracking for interacting with the environment without any additional hand controller and various augmented reality (AR) technologies that allow researchers to incorporate real-world and virtual elements into their studies. One of the key rationales for using VR for naturalistic cognitive tasks is to be able to mimic the environments and functions of everyday life as closely as possible while being able to measure behavior accurately. As pointed out by Slater and Sanchez-Vives (2016), *"VR is different from other forms of human-computer interfaces since the human participates in the virtual world rather than uses it"*, and eventually there will be a paradigm shift with new ways of presenting tasks. Hopefully, these new advancements will be embraced by the research community to develop new task versions with even greater clinical and research utility.

4.9 Conclusion

The current study fills an essential gap in the literature as, to our knowledge, it is the first study to compare FSD and HMD implementations of a naturalistic, open-ended task. This is particularly important, as naturalistic tasks might become the hallmark of VR-based cognition research by taking advantage of the technology's benefits to the fullest (Parsons et al., 2017). Our results show great similarity between the results acquired with the FSD and HMD versions of EPELI, but also distinctive strengths and benefits associated with each version. This information is beneficial not only for the future use of EPELI, but also for researchers developing other naturalistic VR tasks. The feasibility of FSD-EPELI for remote testing also received support. The issue of remote testing is also very timely, as online testing is nowadays common as a cost-effective and flexible alternative to traditional laboratory-based research. We hope that this study will in its part further the naturalistic cognitive research, which has a huge potential to broaden our understanding of human goal-directed behavior.

Data availability statement

In compliance with the research permission by the Ethics Committee of the Helsinki University Hospital, supporting data for this study is not available due to patient confidentiality restrictions. Requests to access the datasets should be directed to erik.seesjarvi@helsinki.fi.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Helsinki University Hospital, Helsinki, Finland. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

JS, ES, and ML designed the experiment. The EPELI task was designed with equal contribution by JS, ES, and ML. ES and KK recruited the participants and collected and preprocessed the data. ES analyzed the data. ES, JS, and ML wrote the manuscript, which was commented and complemented by KK, and agreed on by all authors.

Funding

The study was supported by the Academy of Finland (grants #325981, #328954, and #353518 to JS, grant #323251 to ML). ES received support from the Finnish Cultural Foundation (grant #00201002), the Arvo and Lea Ylppo Foundation (grant #202010005), and the Instrumentarium Science Foundation (grant #200005).

Acknowledgments

We thank Sascha Zuber for reading an earlier version of the manuscript and providing many beneficial comments on it.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frvir.2023.1138240/full#supplementary-material

References

Alapakkam Govindarajan, M. A., Archambault, P. S., and Laplante-El Haili, Y. (2022). Comparing the usability of a virtual reality manual wheelchair simulator in two display conditions. *J. Rehabilitation Assistive Technol. Eng.* 9, 205566832110671. doi:10.1177/20556683211067174

Armstrong, C. M., Reger, G. M., Edwards, J., Rizzo, A. A., Courtney, C. G., and Parsons, T. D. (2013). Validity of the virtual reality Stroop task (VRST) in active duty military. *J. Clin. Exp. Neuropsychology* 35 (2), 113–123. doi:10.1080/13803395.2012.740002

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics. Available at: https://CRAN.R-project.org/package=gridExtra.

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., and Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge neuropsychological test automated battery: A within-subjects counterbalanced study. *J. Med. Internet Res.* 22 (8), e16792. doi:10.2196/16792

Barnett, M. D., Childers, L. G., and Parsons, T. D. (2021). A virtual kitchen protocol to measure everyday memory functioning for meal preparation. *Brain Sci.* 11 (5), 571. doi:10.3390/brainsci11050571

Barrett, R. C. A., Poe, R., O'Camb, J. W., Woodruff, C., Harrison, S. M., Dolguikh, K., et al. (2022). Comparing virtual reality, desktop-based 3D, and 2D versions of a category learning experiment. *PLOS ONE* 17 (10), e0275119. doi:10.1371/journal.pone.0275119

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67 (1). doi:10.18637/jss.v067.i01

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., and Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of clinical Neuropsychology and the national Academy of Neuropsychology. *Archives Clin. Neuropsychology* 27 (3), 362–373. doi:10.1093/arclin/acs027

Ben-Shachar, M., Lüdecke, D., and Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *J. Open Source Softw.* 5 (56), 2815. doi:10.21105/joss.02815

Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nat. Rev. Neurosci.* 12 (12), 752–762. doi:10.1038/nrn3122

Brooks, J., Lodge, R., and White, D. (2017). Comparison of a head-mounted display and flat screen display during a micro-UAV target detection task. *Proc. Hum. Factors Ergonomics Soc. Annu. Meet.* 61 (1), 1514–1518. doi:10.1177/1541931213601863

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., Coates, M-A. L., Dawson, D. R., et al. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *J. Int. Neuropsychological Soc.* 12 (2), 194–209. doi:10.1017/S1355617706060310

Campbell, Z., Zakzanis, K. K., Jovanovski, D., Joordens, S., Mraz, R., and Graham, S. J. (2009). Utilizing virtual reality to improve the ecological validity of clinical Neuropsychology: An fMRI case study elucidating the neural basis of planning by comparing the tower of london with a three-dimensional navigation task. *Appl. Neuropsychol.* 16 (4), 295–306. doi:10.1080/09084280903297891

Caroux, L. (2023). Presence in video games: A systematic review and meta-analysis of the effects of game design choices. *Appl. Ergon.* 107, 103936. doi:10.1016/j.apergo.2022.103936

Chan, R., Shum, D., Toulopoulou, T., and Chen, E. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives Clin. Neuropsychology* 23 (2), 201–216. doi:10.1016/j.acn.2007.08.010

Chang, C. W., Li, M., Yeh, S. C., Chen, Y., and Rizzo, A. (2020). Examining the effects of HMDs/FSDs and gender differences on cognitive processing ability and user experience of the Stroop task-embedded virtual reality driving system (STEVRDS). *IEEE Access* 8, 69566–69578. doi:10.1109/access.2020.2966564

Chicchi Giglioli, I. A., Pérez Gálvez, B., Gil Granados, A., and Alcañiz Raya, M. (2021). 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, Behav. Soc. Netw.* 2020, 0560. doi:10.1089/cyber.2020.0560

Cipresso, P., Albani, G., Serino, S., Pedroli, E., Pallavicini, F., Mauro, A., et al. (2014). Virtual multiple errands test (VMET): A virtual reality-based tool to detect early executive functions deficit in Parkinson's disease. *Front. Behav. Neurosci.* 8, 405. doi:10.3389/fnbeh.2014.00405

Cipresso, P., Giglioli, I. A. C., Raya, M. A., and Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Front. Psychol.* 9, 2086. doi:10.3389/fpsyg.2018.02086

Conner, N. O., Freeman, H. R., Jones, J. A., Luczak, T., Carruth, D., Knight, A. C., et al. (2022). Virtual reality induced symptoms and effects: Concerns, causes, assessment & mitigation. *Virtual Worlds* 1 (2), 130–146. doi:10.3390/virtualworlds1020008

Constantinou, M., Bauer, L., Ashendorf, L., Fisher, J., and Mccaffrey, R. (2005). Is poor performance on recognition memory effort measures indicative of generalized poor performance on neuropsychological tests? *Archives Clin. Neuropsychology* 20 (2), 191–198. doi:10.1016/j.acn.2004.06.002

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS ONE* 8 (3), e57410. doi:10.1371/journal.pone.0057410

Delis, D. C., Kramer, J. H., Kaplan, E., and Holdnack, J. (2004). Reliability and validity of the delis-kaplan executive function system: An update. *J. Int. Neuropsychological Soc.* 10 (2), 301–303. doi:10.1017/S1355617704102191

Di Natale, A. F., Repetto, C., Riva, G., and Villani, D. (2020). Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research. *Br. J. Educ. Technol.* 51 (6), 2006–2033. doi:10.1111/bjet.13030

Díaz-Orueta, U., Garcia-López, C., Crespo-Eguílaz, N., Sánchez-Carpintero, R., Climent, G., and Narbona, J. (2014). AULA virtual reality test as an attention measure: Convergent validity with Conners' Continuous Performance Test. *Child. Neuropsychol.* 20 (3), 328–342. doi:10.1080/09297049.2013.792332

Dowle, M., and Srinivasan, A. (2021). data.table: Extension of `data.frame. Available at: https://CRAN.R-project.org/package=data.table.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives Clin. Neuropsychology* 27 (3), 248–261. doi:10.1093/arclin/acr120

Feenstra, H. E. M., Vermeulen, I. E., Murre, J. M. J., and Schagen, S. B. (2017). Online cognition: Factors facilitating reliable online neuropsychological test results. *Clin. Neuropsychologist* 31 (1), 59–84. doi:10.1080/13854046.2016.1190405

Gagolewski, M. (2020). R package stringi: Character string processing facilities. Available at: http://www.gagolewski.com/software/stringi/.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bull. Rev.* 19 (5), 847–857. doi:10.3758/s13423-012-0296-9

Gioia, G. A., Isquith, P. K., Guy, S. C., and Kenworthy, L. (2000). *Behavior rating inventory of executive function: Brief*. Odessa, FL: Psychological Assessment Resources.

Hatfield, G. (2002). Psychology, philosophy, and cognitive science: Reflections on the history and philosophy of experimental Psychology. *Mind Lang.* 17 (3), 207–232. doi:10.1111/1468-0017.00196

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., and Hart, R. P. (2010). Official position of the American Academy of clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *Clin. Neuropsychologist* 24 (8), 1267–1278. doi:10.1080/13854046.2010.526785

Jovanovski, D., Zakzanis, K., Campbell, Z., Erb, S., and Nussbaum, D. (2012). Development of a Novel, ecologically oriented virtual reality measure of executive function: The multitasking in the city test. *Appl. Neuropsychol. Adult* 19 (3), 171–182. doi:10.1080/09084282.2011.643955

Jylkkä, J., Ritakallio, L., Merzon, L., Kangas, S., Kliegel, M., Zuber, S., et al. (2023). Assessment of goal-directed behavior with the 3D videogame EPELI: Psychometric features in a web-based adult sample. *PLOS ONE* 18 (3), e0280717. doi:10.1371/journal.pone.0280717

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3 (3), 203–220. doi:10.1207/s15327108ijap0303_3

Khan, A., Sharma, N. K., and Dixit, S. (2008). Cognitive load and task condition in event- and time-based prospective memory: An experimental investigation. *J. Psychol.* 142 (5), 517–532. doi:10.3200/JRLP.142.5.517-532

Kim, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Commun. Stat. Appl. Methods* 22 (6), 665–674. doi:10.5351/CSAM.2015.22.6.665

Kothgassner, O. D., and Felnhofer, A. (2020). Does virtual reality help to cut the Gordian knot between ecological validity and experimental control? *Ann. Int. Commun. Assoc.* 44 (3), 210–218. doi:10.1080/23808985.2020.1792790

Kourtesis, P., Collina, S., Doumas, L. A. A., and MacPherson, S. E. (2021). Validation of the virtual reality everyday assessment lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity. *J. Int. Neuropsychological Soc.* 27 (2), 181–196. doi:10.1017/S1355617720000764

Kourtesis, P., Collina, S., Doumas, L. A. A., and MacPherson, S. E. (2019). Validation of the virtual reality neuroscience questionnaire: Maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Front. Hum. Neurosci.* 13, 417. doi:10.3389/fnhum.2019.00417

Kourtesis, P., Korre, D., Collina, S., Doumas, L. A. A., and MacPherson, S. E. (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. *Front. Comput. Sci.* 1, 12. doi:10.3389/fcomp.2019.00012

Kourtesis, P., and MacPherson, S. E. (2021). How immersive virtual reality methods may meet the criteria of the national Academy of Neuropsychology and American Academy of clinical Neuropsychology: A software review of the virtual reality everyday assessment lab (VR-EAL). *Comput. Hum. Behav. Rep.* 4, 100151. doi:10.1016/j.chbr.2021.100151

Krohn, S., Tromp, J., Quinque, E. M., Belger, J., Klotzsche, F., Rekers, S., et al. (2020). Multidimensional evaluation of virtual reality paradigms in clinical Neuropsychology: Application of the VR-check framework. *J. Med. Internet Res.* 22 (4), e16724. doi:10.2196/16724

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *J. Stat. Softw.* 82 (13). doi:10.18637/jss.v082.i13

Li, G., Anguera, J. A., Javed, S. V., Khan, M. A., Wang, G., and Gazzaley, A. (2020). Enhanced attention using head-mounted virtual reality. *J. Cognitive Neurosci.* 32 (8), 1438–1454. doi:10.1162/jocn_a_01560

Makransky, G., Terkildsen, T. S., and Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learn. Instr.* 60, 225–236. doi:10.1016/j.learninstruc.2017.12.007

Martel, E., Su, F., Gerroir, J., Hassan, A., Girouard, A., and Muldner, K. (2015). Diving head-first into virtual reality: Evaluating HMD control schemes for VR games. *FDG* 1–5.

Merzon, L., Pettersson, K., Aronen, E. T., Huhdanpää, H., Seesjärvi, E., Henriksson, L., et al. (2022). Eye movement behavior in a real-world virtual reality task reveals ADHD in children. *Sci. Rep.* 12 (1), 20308. doi:10.1038/s41598-022-24552-4

Mioni, G., Rendell, P. G., Stablum, F., Gamberini, L., and Bisiacchi, P. S. (2015). Test–retest consistency of virtual week: A task to investigate prospective memory. *Neuropsychol. Rehabil.* 25 (3), 419–447. doi:10.1080/09602011.2014.941295

Mullen, G., and Davidenko, N. (2021). Time compression in virtual reality. *Timing & Time Percept.* 9 (4), 377–392. doi:10.1163/22134468-bja10034

Neguț, A., Matu, S.-A., Sava, F. A., and David, D. (2016). Virtual reality measures in neuropsychological assessment: A meta-analytic review. *Clin. Neuropsychologist* 30 (2), 165–184. doi:10.1080/13854046.2016.1144793

Ouellet, É., Boller, B., Corriveau-Lecavalier, N., Cloutier, S., and Belleville, S. (2018). The virtual shop: A new immersive virtual reality environment and scenario for the assessment of everyday memory. *J. Neurosci. Methods* 303, 126–135. doi:10.1016/j.jneumeth.2018.03.010

Pallavicini, F., Ferrari, A., Pepe, A., Garcea, G., Zanacchi, A., and Mantovani, F. (2018). "Effectiveness of virtual reality survival horror games for the emotional elicitation: Preliminary insights using resident evil 7: Biohazard," in *Universal access in human-computer interaction. Virtual, augmented, and intelligent environments*. Editors M. Antona and C. Stephanidis (New York City: Springer International Publishing). doi:10.1007/978-3-319-92052-8_8

Pallavicini, F., and Pepe, A. (2019). Comparing player experience in video games played in virtual reality or on desktop displays: Immersion, flow, and positive emotions. *Ext. Abstr. Annu. Symposium Computer-Human Interact. Play Companion Ext. Abstr.* 2019, 195–210. doi:10.1145/3341215.3355736

Pallavicini, F., Pepe, A., and Minissi, M. E. (2019). Gaming in virtual reality: What changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simul. Gaming* 50 (2), 136–159. doi:10.1177/1046878119831420

Palmisano, S., Allison, R. S., and Kim, J. (2020). Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose. *Front. Virtual Real.* 1, 587698. doi:10.3389/frvir.2020.587698

Pan, X., and Hamilton, A. F. de C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *Br. J. Psychol.* 109 (3), 395–417. doi:10.1111/bjop.12290

Parsons, T. D., and Barnett, M. (2017). Validity of a newly developed measure of memory: Feasibility study of the virtual environment grocery store. *J. Alzheimer's Dis.* 59 (4), 1227–1235. doi:10.3233/JAD-170295

Parsons, T. D., Carlew, A. R., Magtoto, J., and Stonecipher, K. (2017). The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical Neuropsychology. *Neuropsychol. Rehabil.* 27 (5), 777–807. doi:10.1080/09602011.2015.1109524

Parsons, T. D., and Rizzo, A. "Skip" (2019). "A review of virtual classroom environments for neuropsychological assessment," in *Virtual reality for psychological and neurocognitive interventions*. Editors A. Skip" Rizzo and S. Bouchard (New York: Springer), 247–265. doi:10.1007/978-1-4939-9482-3_11

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Front. Hum. Neurosci.* 9, 660. doi:10.3389/fnhum.2015.00660

Pedersen, T. L. (2020). patchwork: The composer of plots. Available at: https://CRAN.R-project.org/package=patchwork.

Pieri, L., Tosi, G., and Romano, D. (2023). Virtual reality technology in neuropsychological testing: A systematic review. *J. Neuropsychology* 55, 12304. doi:10.1111/jnp.12304

Porffy, L. A., Mehta, M. A., Patchitt, J., Boussebaa, C., Brett, J., D'Oliveira, T., et al. (2022). A Novel virtual reality assessment of functional cognition: Validation study. *J. Med. Internet Res.* 24 (1), e27641. doi:10.2196/27641

Rabbitt, P. (Editor) (2004). "Introduction: Methodologies and models in the study of executive function," *Methodology of frontal and executive function*. 1st ed. (England, UK: Routledge), 9–45. doi:10.4324/9780203344187-5

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Rand, D., Rukan, S. B.-A., Weiss, P. L., and Katz, N. (2009). Validation of the Virtual MET as an assessment tool for executive functions. *Neuropsychol. Rehabil.* 19 (4), 583–602. doi:10.1080/09602010802469074

Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Pedroli, E., Cipresso, P., et al. (2012). Validating the neuro VR-based virtual version of the multiple errands test: Preliminary results. *Presence Teleoperators Virtual Environ.* 21 (1), 31–42. doi:10.1162/PRES_a_00077

Rendell, P. G., and Craik, F. I. M. (2000). Virtual week and actual week: Age-related differences in prospective memory. *Appl. Cogn. Psychol.* 14 (7), S43–S62. doi:10.1002/acp.770

Repovš, G., and Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive Psychology. *Neuroscience* 139 (1), 5–21. doi:10.1016/j.neuroscience.2005.12.061

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research*. United States: Northwestern University.

Rotenberg, S., Ruthralingam, M., Hnatiw, B., Neufeld, K., Yuzwa, K. E., Arbel, I., et al. (2020). Measurement properties of the multiple errands test: A systematic review. *Archives Phys. Med. Rehabilitation* 101 (9), 1628–1642. doi:10.1016/j.apmr.2020.01.019

Ruse, S. A., Harvey, P. D., Davis, V. G., Atkins, A. S., Fox, K. H., and Keefe, R. S. E. (2014). Virtual reality functional capacity assessment in schizophrenia: Preliminary data regarding feasibility and correlations with cognitive and functional capacity performance. *Schizophrenia Res. Cognition* 1 (1), e21–e26. doi:10.1016/j.scog.2014.01.004

Seesjärvi, E., Puhakka, J., Aronen, E. T., Hering, A., Zuber, S., Merzon, L., et al. (2022a). Epeli: A novel virtual reality task for the assessment of goal-directed behavior in real-life contexts. *Psychol. Res.* 22, 1770. doi:10.1007/s00426-022-01770-z

Seesjärvi, E., Puhakka, J., Aronen, E. T., Lipsanen, J., Mannerkoski, M., Hering, A., et al. (2022b). Quantifying ADHD symptoms in open-ended everyday life contexts with a new virtual reality task. *J. Atten. Disord.* 26 (11), 1394–1411. doi:10.1177/10870547211044214

Shallice, T., and Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. *Brain* 114 (2), 727–741. doi:10.1093/brain/114.2.727

Shelstad, W. J., Smith, D. C., and Chaparro, B. S. (2017). Gaming on the rift: How virtual reality affects game user satisfaction. *Proc. Hum. Factors Ergonomics Soc. Annu. Meet.* 61 (1), 2072–2076. doi:10.1177/1541931213602001

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *Br. J. Psychol.* 109 (3), 431–433. doi:10.1111/bjop.12305

Slater, M., and Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Front. Robotics AI* 3, 3. doi:10.3389/frobt.2016.00074

Stevens, A., Friedel, E., Mehren, G., and Merten, T. (2008). Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. *Psychiatry Res.* 157 (1–3), 191–200. doi:10.1016/j.psychres.2007.01.003

Suh, A., and Prophet, J. (2018). The state of immersive technology research: A literature analysis. *Comput. Hum. Behav.* 86, 77–90. doi:10.1016/j.chb.2018.04.019

Tan, C. T., Leong, T. W., Shen, S., Dubravs, C., and Si, C. (2015). "Exploring gameplay experiences on the Oculus rift," in *Proceedings of the 2015 annual symposium on computer-human interaction in play* (United States: Association for Computing Machinery), 253–263. doi:10.1145/2793107.2793117

Teixeira, J., and Palmisano, S. (2021). Effects of dynamic field-of-view restriction on cybersickness and presence in HMD-based virtual reality. *Virtual Real.* 25 (2), 433–445. doi:10.1007/s10055-020-00466-2

Ventura, S., Brivio, E., Riva, G., and Baños, R. M. (2019). Immersive versus non-immersive experience: Exploring the feasibility of memory assessment through 360° technology. *Front. Psychol.* 10, 2509. doi:10.3389/fpsyg.2019.02509

Walch, M., Frommel, J., Rogers, K., Schüssel, F., Hock, P., Dobbelstein, D., et al. (2017). "Evaluating VR driving simulation from a player experience perspective," in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, May 6 - 11, 2017, Denver Colorado USA, 2982–2989. doi:10.1145/3027063.3053202

Webb, S. S., Jespersen, A., Chiu, E. G., Payne, F., Basting, R., Duta, M. D., et al. (2021). The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test. *Neuropsychol. Rehabil.* 32, 1007–1032. doi:10.1080/09602011.2020.1862679

Wechsler, D. (1997). *WAIS-III: Wechsler adult intelligence scale*. 3rd ed. San Antonio, TX: Psychological Corporation.

Weech, S., Kenny, S., and Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Front. Psychol.* 10, 158. doi:10.3389/fpsyg.2019.00158

West, L. K., Curtis, K. L., Greve, K. W., and Bianchini, K. J. (2011). Memory in traumatic brain injury: The effects of injury severity and effort on the wechsler memory scale-III: WMS-III and TBI. *J. Neuropsychology* 5 (1), 114–125. doi:10.1348/174866410X521434

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4 (43), 1686. doi:10.21105/joss.01686

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. Available at: https://CRAN.R-project.org/package=stringr.

Wickham, H., François, R., Henry, L., and Müller, K. (2021). dplyr: A Grammar of Data Manipulation. Available at: https://CRAN.R-project.org/package=dplyr

Witmer, B. G., Jerome, C. J., and Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence Teleoperators Virtual Environ.* 14 (3), 298–312. doi:10.1162/105474605323384654

Woods, S., Delis, D., Scott, J., Kramer, J., and Holdnack, J. (2006). The California verbal learning test – second edition: Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives Clin. Neuropsychology* 21 (5), 413–420. doi:10.1016/j.acn.2006.06.002

World Health Organization (1992). "World Health organization," in *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (Geneva, Switzerland: World Health Organization).

Yao, S., and Kim, G. (2019). "The effects of immersion in a virtual reality game: Presence and physical activity," in *HCI in games*. Editor X. Fang (New York City: Springer International Publishing). doi:10.1007/978-3-030-22602-2_18

Zuber, S., Haas, M., Framorando, D., Ballhausen, N., Gillioz, E., Künzi, M., et al. (2021). The Geneva space cruiser: A fully self-administered online tool to assess prospective memory across the adult lifespan. *Memory* 1, 117–132. doi:10.1080/09658211.2021.1995435