University of Helsinki
Dissertationes Universitatis Helsingiensis 41/2024

Department of Psychology and Logopedics

Doctoral Programme in Clinical Research, Faculty of Medicine,
University of Helsinki, Finland

Function-led Assessment of Children's Goal-directed Behavior and ADHD Symptoms in Virtual Reality

Erik Seesjärvi

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in hall 2, Haartman institute, on the 1st of March, 2024, at 12 o'clock.

Helsinki 2024

Thesis supervisors

Juha Salmitaival, Docent, Academy Fellow
Department of Neuroscience and Biomedical Engineering
Aalto University, Espoo, Finland

Matti Laine, Professor
Department of Psychology
Åbo Akademi University, Turku, Finland

Eeva T. Aronen, Professor
University of Helsinki and Helsinki University Hospital, Child Psychiatry and New Children's Hospital, Pediatric Research Center, Helsinki, Finland

Pre-examiners

Giuseppe Riva, Professor
Department of Psychology, Università Cattolica del Sacro Cuore, Italy

Alexandra Voinescu, PhD
Department of Psychology, The University of Bath, the United Kingdom

Opponent

Claudia Repetto, Associate Professor
Department of Psychology, Università Cattolica del Sacro Cuore, Milan, Italy

The Faculty of Medicine uses the Ouriginal system (plagiarism recognition) to examine all doctoral dissertations.

Publisher: University of Helsinki
Series: Dissertationes Universitatis Helsingiensis 41/2024

ISBN 978-951-51-9647-7 (print)
ISBN 978-951-51-9648-4 (online)
ISSN 2954-2898 (print)
ISSN 2954-2952 (online)
PunaMusta, Joensuu 2024

Abstract

Objective, reliable, and ecologically valid measurement of goal-directed behavior and related cognitive processes, such as executive functions and prospective memory, has proven to be challenging. Difficulties in these cognitive domains can have severe consequences for everyday life, but current neuropsychological tests may not be optimal tools for the comprehensive assessment of such problems. It has been suggested that naturalistic tasks that simulate everyday life activities could provide the researcher and clinician with complementary means to better evaluate these important domains while allowing the assessment of other aspects of behavior, such as the symptoms of various clinical disorders like attention deficit/hyperactivity disorder (ADHD).

The main aim of this Dissertation was to develop and apply a new virtual reality (VR) task, named Executive Performance in Everyday LIving (EPELI), as a more ecologically valid alternative for the assessment of goal-directed behavior in children. More specific aims included examining EPELI's ecological validity, discriminating capacity for ADHD, reliability, usability, and potential cybersickness symptoms. Also, eye movement behavior was quantified to study visual attention in ADHD. In addition to the immersive head-mounted display version, a non-immersive version for flat screen displays was developed and tested. A total of 85 children with ADHD and 146 typically developing children aged 9 to 13 years took part in the four studies comprising this Dissertation.

The results revealed that EPELI shows ecological validity by being associated with parent-rated everyday problems of attention and executive function. Furthermore, task performance in EPELI successfully discriminates between children with and without ADHD. Eye movement behavior can be used to further improve this discriminating capacity and to quantify visual attention in greater detail. EPELI has good usability, is considered enjoyable by children, and does not cause cybersickness symptoms. Moreover, most of its measures show adequate reliability. The non-immersive flat screen display version can be used remotely with commonly available hardware, but children consider it less immersive than the head-mounted display version.

All in all, EPELI was shown to be a valuable complement to the assessment of goal-directed behavior in children. To our knowledge, it is the first immersive VR task for school-aged children that can be used to quantify goal-directed behavior and ADHD symptoms in open-ended everyday scenarios. EPELI provides rich yet well-controlled objective data that reflect these behaviors and symptoms. In clinical settings, such data could be used to complement survey instruments and interviews, which are subjective. Being able to replicate some aspects of real-life problems in simulated settings could also facilitate communication with the child, caregivers, and other stakeholders such as teachers. These findings encourage to extend the research on VR with function-led tasks like EPELI to other clinical neuropsychiatric conditions and further simulated contexts.

Finnish summary

Tavoitesuuntautuneen käyttäytymisen ja sen taustalla vaikuttavien toiminnanohjaus- ja muistiprosessien objektiivinen, luotettava ja ekologisesti validi arviointi on osoittautunut haasteelliseksi. Vaikka näiden tiedonkäsittelyprosessien vaikeudet voivat merkittävästi heikentää arjen toimintakykyä, nykyiset neuropsykologiset testit eivät välttämättä tarjoa riittäviä välineitä niiden kattavaan arviointiin. On ehdotettu, että arkisia toimintoja ja tilanteita jäljittelevät tehtävät voisivat täydentää nykyisiä arviointimenetelmiä mahdollistaen samalla aktiivisuuden ja tarkkaavuuden häiriön (ADHD) oireiden ja myös muiden käyttäytymispiirteiden arvioimisen.

Tämän väitöskirjan päätavoite oli kehittää uusi, virtuaalitodellisuutta (*engl.* Virtual Reality, VR) hyödyntävä ja ekologisesti validi lasten tavoitesuuntautuneen toiminnan arviointimenetelmä, EPELI (*engl.* Executive Performance in Everyday LIving). Alatavoitteina oli tarkastella EPELIn ekologista validiteettia, erottelukykyä suhteessa ADHD:hen, reliabiliteettia, käytettävyyttä, mahdollisia pahoinvointioireita ja silmänliikkeiden käyttöä näönvaraisen tarkkaavuuden arvioinnissa. EPELIstä kehitettiin VR-laseilla käytettävän version lisäksi tavallisille tietokonenäytöille soveltuva versio. Tutkimukseen osallistui yhteensä 85 ADHD-diagnoosin saanutta ja 146 tavanomaisesti kehittynyttä 9–13-vuotiasta lasta.

Tulokset osoittivat, että EPELI on yhteydessä vanhemman arvioon lapsen toiminnanohjausvaikeuksista, mikä tukee EPELIn ekologista validiteettia. Suoriutuminen EPELIssä erottelee ADHD-diagnoosin saaneita ja tyypillisesti kehittyviä lapsia. Silmänliikkeitä voidaan hyödyntää erottelutarkkuuden parantamiseen ja näönvaraisen tarkkaavuuden tarkempaan mittaamiseen. Menetelmän käytettävyys on hyvä, lapset nauttivat sen suorittamisesta, eikä sen käyttö aiheuta pahoinvointia. Reliabiliteetti on riittävä suurimmassa osassa EPELIn mittareita. Tavallisille tietokonenäytöille soveltavaa versiota voidaan käyttää myös etäyhteyden välityksellä, mutta lapset pitävät sitä vähemmän immersiivisenä eli mukaansatempaavana kuin VR-lasiversiota.

Kokonaisuutena tutkimus osoitti, että EPELI on hyödyllinen lisä lasten tavoitesuuntautuneen toiminnan arviointiin. Tietojemme mukaan EPELI on ensimmäinen VR-laseilla toteutettava tehtävä, jolla voidaan mitata lasten tavoitesuuntautunutta toimintaa ja ADHD-oireita avoimessa, luonnollisenkaltaisessa tehtäväympäristössä. Menetelmä tuottaa monipuolista, tarkasti kontrolloitua ja objektiivista tietoa tavoitesuuntautuneesta käyttäytymisestä ja ADHD-oireista. Kliinisessä käytössä tällaista tietoa voitaisiin hyödyntää subjektiivisten arviointimenetelmien kuten kyselylomakkeiden ja haastatteluiden rinnalla. EPELIä voidaan mahdollisesti myös hyödyntää tavoitesuuntautuneen toiminnan vaikeuksien havainnollistamiseen ja niistä keskustelemiseen. Tutkimuslöydökset kannustavat jatkamaan VR-pohjaisten, luonnollisia tilanteita jäljittelevien tehtävien kehittämistä muun muassa muiden neuropsykiatristen sairauksien oireiden tutkimisessa.

Preface

> "[T]he argument here is that one should consider in the design of new tasks the demands that the "real world" may make but that are not presented to the participant performing *existing* tests in a lab setting. By not considering these aspects, we may be missing decrements in many aspects of cognition which are critical to competence in everyday life." (Burgess et al., 2006)

My inspiration to get involved in developing a novel neuropsychological task was grounded in observations made during clinical work. While certain aspects of children's cognition, such as logical reasoning and language, appeared to yield themselves to be evaluated in the structured one-to-one sessions of neuropsychological assessment, others surely seemed to elude trustworthy evaluation under such circumstances. For the aspiring neuropsychologist-to-be, interviews and rating measures involving the child, parents, and other adults seemed to provide an important alternative to the performance-based measures. Yet, they occasionally appeared to mostly reflect the subjective attitudes, motivations, and expectations of the assessor, not the skills or difficulties of the assessee. Very importantly, the psychologist's own observations and interpretations are also subjective. Thus, becoming overconfident in one's own skills, honed by the accumulating clinical experience, seemed like a bad strategy to overcome the limitations of tools currently available to the clinical neuropsychologist. Luckily, there was already an abundant body of literature on how naturalistic tasks may improve the assessment of goal-directed behavior. Moreover, recent improvements in VR technology supported the idea that it was a good time to have a go on developing a new naturalistic VR task for children. Despite facing some setbacks and challenges, like performing the data collection in 2020–2021 amidst the fluctuating restrictions caused by the COVID-19 pandemic, I think we managed to pull this through surprisingly well. For this, THANK YOU for all the wonderful people involved.

First, I want to thank my outstanding supervisors for doing their part with utmost professionalism and dedication. Juha, thank for always keeping up with what has been agreed on and pursuing research with such an inspirational drive. I admire your superhuman efficiency in pursuing goal-directed research. Thank you also for the support during moments when things were not that easy and coming

up with solutions to all imaginable and unimaginable problems along the way. And btw, sorry for the skiing. Matti, I am deeply grateful that you agreed to supervise me and share your truly enormous experience of neuropsychology with me. To have been able to develop and discuss all the ideas with you has really been most crucial for me for believing in what we were doing here. As a bonus, you are by far the nicest grand old man of neuropsychology in Finland. This has been confirmed by numerous independent raters. Eeva, your support and guidance has been essential to this project. Furthermore, I have learned a lot from you about ADHD, related psychiatric conditions, and clinical interviewing. As a clinical neuropsychologist, I really value this.

This work would not have been possible without the support of Peili Vision company that implemented the EPELI versions used here. Thank you, Topi, Mikko, Joonas, Vesa, Jussi, Akviliina, and the others! I really appreciate a startup that takes the challenge of providing new solutions to the health care sector. Because, if we do not have the tools, we just do not have the tools. It's simple as that and everyone devoted to providing better care for the patients and their families should acknowledge this.

I am grateful for all the coauthors for their hard work and effort. Here, a special mention needs to be given to our collaborators Matthias Kliegel, Alexandra Hering, and Sascha Zuber, whose world-leading expertise on prospective memory has been crucial. Liya, I am very happy that you choose to join Juha's project and brought your own expertise to it. Without you, things simply would not be where they are today. Jasmin and Kaisla, big thank you for taking part in the project and especially in the data collection. Working with kids is often very rewarding, yet it can also be challenging from time to time. I believe you both have a natural talent for this. Furthermore, warm thanks to all the following people who have supported my work on this Thesis: Jussi Jylkkä, Liisa Ritakallio, Veli-Matti Saarinen, Jaakko Kauramäki, Jari Lipsanen, Minna Mannerkoski, Hanna Huhdanpää, Kati Pettersson, Linda Henriksson, Joe MacInnes, Emiliano Macaluso, Tessa Tolonen, Sofia Tauriainen, Daniel Fellman, Daniel Wärnå, and Juuso Räsänen. Eeva Eskola, thank you for taking the time to read the thesis summary draft and providing many great insights. I also want to thank Kenneth Quek for proofreading the manuscripts of Studies I, III, and IV and taking his time to answer all my silly questions about the English language. I apologize for all new grammatical errors I've introduced to the text(s) during the review processes that took place after your proofreading.

This Dissertation was funded by grants from the Finnish Cultural Foundation (#00190963 and #00201002), the Arvo and Lea Ylppo Foundation (#202010005), and the Instrumentarium Science Foundation (#200005). I am thankful for this essential financial aid and the support of Päivi Klemola, Irmeli Kosonen, and all the others who have helped me here.

I have done this Thesis while working as a clinical neuropsychologist in Helsinki University Hospital. The support and ideas from neuropsychologists, both seniors and newcomers, have been invaluable. I extend my gratitude to Susanna Huju, Henri Lehtinen, Satu Korpela, and all the others involved. I also want to thank my thesis supervision committee members, Leena Haataja and Vesa Närhi, for their wise advises regarding this research and how to continue pursuing goals that seem meaningful and important.

The recruitment of participants has included numerous people from several schools in Kirkkonummi and Espoo, Helsinki University Hospital, Finnish ADHD Foundation (including subassociations Aisti, Neuris, and Itua), the Espoo City Child Psychiatric Unit, the Vantaa Family Counselling Unit, and ProNeuron LTD. This includes many who have done their work pro bono. For this, I am very thankful to Annamaija Kylä-Setälä, Anu Virtanen, Hanna Uurainen, Jari Hämäläinen, Juulia Paavonen, Katri Lahti, Anu Kivistö, Henna Kainomaa, Marja Kantele, Sari Korpirinne, Jaana Wessman, Marjo Raita, Päivi Harjula, Anssi Iivonen, Mikko Hommo, Eeva Lumiaro, Jukka Sarpila, Arja Rantala, all the wonderful teachers, all the persons I forgot, and all of those that I do not even know.

EPELI would not exist without the children and families who gave their time to take part in the studies and piloting that commenced before Study I. It means so much for me that you agreed to do so. Furthermore, the discussions during the assessments and parent feedback sessions have been truly inspirational. You know things that we, as researchers and clinicians, do not. By sharing your own experiences and perspectives, you really help us to do better.

Finally, I want to thank my wife, children, parents, parents-in-law, siblings, other family members, and close friends for your support. This PhD would simply not exist without it. Thank you for what matters most beyond life.

Index

List of original publications

This Dissertation is based on the following publications:

I Seesjärvi, E., Puhakka, J., Aronen, E. T., Lipsanen, J., Mannerkoski, M., Hering, A., Zuber, S., Kliegel, M., Laine, M., & Salmi, J. (2022). Quantifying ADHD symptoms in open-ended everyday life contexts with a new virtual reality task. *Journal of Attention Disorders*, *26*(11), 1394–1411. https://doi.org/10.1177/10870547211044214

II Merzon, L., Pettersson, K., Aronen, E. T., Huhdanpää, H., Seesjärvi, E., Henriksson, L., MacInnes, W. J., Mannerkoski, M., Macaluso, E., & Salmi, J. (2022). Eye movement behavior in a real-world virtual reality task reveals ADHD in children. *Scientific Reports*, *12*, 20308. https://doi.org/10.1038/s41598-022-24552-4

III Seesjärvi, E., Puhakka, J., Aronen, E. T., Hering, A., Zuber, S., Merzon, L., Kliegel, M., Laine, M., & Salmi, J. (2023). EPELI: A novel virtual reality task for the assessment of goal-directed behavior in real-life contexts. *Psychological Research, 87*, 1899–1916. https://doi.org/10.1007/s00426-022-01770-z

IV Seesjärvi, E., Laine, M., Kasteenpohja, K., & 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. *Frontiers in Virtual Reality*, 4:1138240. http://doi.org/10.3389/frvir.2023.1138240

The publications are referred to in the text by their roman numerals. The articles are reprinted with the kind permission of the copyright holders.

List of abbreviations

ADHD	Attention-Deficit/Hyperactivity Disorder
ADHD-RS	the ADHD Rating Scale-IV
AIC	Akaike Information Criterion
ANOVA	analysis of variance
AUC	area under the curve
BIC	the Bayesian information criterion
BRI	behavioral regulation index
BRIEF	the Behavior Rating Inventory of Executive Function
CBCL	the Child Behavior Checklist
CPT	the Continuous Performance Task
EBPM	event-based prospective memory
EF	executive functions
EPELI	Executive Performance in Everyday LIving
EQELI	Executive Questionnaire of Everyday LIfe
F&C	Frogs and Cherries task
FDR	false discovery rate
FOV	field-of-view
FSD	flat screen display
GEC	general executive composite
HEXE	Heidelberger Exekutivfunktionsdiagnostikum Task
HMD	head-mounted display
ICC	intraclass correlation
LMM	linear mixed models
MET	the Multiple Errands Test
MI	metacognition index
NSS	normalized scanpath saliency
PM	prospective memory
ROC	receiver operating characteristic
SRT	Simple Reaction Task
SVM	support vector machine
TBPM	time-based prospective memory
VR	virtual reality

1 Introduction

1.1 Goal-directed behavior and related concepts

The present Dissertation focuses on the assessment of *goal-directed behavior*, that is, behavior taken toward attaining a specific goal (American Psychological Association, 2015). These goals are presented in mind as imagined future states, and they activate and are influenced by other mental content such as knowledge, beliefs, norms, values, and preferences (Doebel, 2020). The higher-level psychological processes that allow goal-directed behaviors are often conceptualized as *executive functions* (EF; Barkley, 2012; Goldstein et al., 2014). EF are needed particularly when dealing with novel tasks, where the individual cannot rely only on previous, well-learned behaviors to reach their goals (see Rabbit, 2004). Thus, an intimate link between goal-directed behavior and EF can be postulated (see Friedman & Robbins, 2022), even though a universally agreed definition of EF has not yet been reached (Barkley, 2012; Goldstein et al., 2014; Jurado & Rosselli, 2007; cf. Diamond, 2013). *Prospective memory* (PM), the memory for to-be-performed future activities (Einstein & McDaniel, 1990; Kliegel et al., 2008b), can also be regarded as a form of goal-directed behavior. As both EF and PM play prominent roles in how goal-directed behavior has been conceptualized and evaluated, a brief overview of these concepts is given next, before moving on to their assessment.

1.1.1 Executive functions

EF is an umbrella term used for a diverse set of hypothesized cognitive processes, including planning, working memory, attention, inhibition, self-monitoring, self-regulation, and initiation (Goldstein et al., 2014; see also Baggetta & Alexander, 2016; Barkley, 2012; Diamond, 2013; Zelazo et al., 2008). EFs relate to multiple important life domains, ranging from success in school to physical health (e.g., Barkley, 2012; Diamond, 2013; Goldstein et al., 2014), and EF deficits are linked to various clinical outcomes (e.g., Pennington and Ozonoff, 1996; Snyder et al., 2015). First EF skills are observable in very early childhood and their maturation is critical from a developmental perspective, as it is related to the development of

autonomy, socioemotional functioning, and academic performance (Best & Miller, 2010; Denham et al., 2015; Diamond and Ling, 2016; Riggs et al., 2006).

Over 30 different EF definitions originating from various empirical findings and theoretical backgrounds have been presented (Barkley, 2012; Goldstein et al., 2014; see also Baggetta & Alexander, 2016). In some models and definitions, EF are conceptualized as purely cognitive processes or constructs without accounting for the impact of emotional processes on cognition (e.g., Miyake et al., 2000), whereas other authors try to also embed these processes in their models (e.g., Barkley, 2012). Recently, Doebel (2020) suggested that instead of defining EF as a set of separate domain-general cognitive processes, EF should be viewed as employment of control when pursuing goals that activate and are influenced by diverse mental content such as knowledge, beliefs, norms, values, and preferences. Although the first publications that used the term EF located its neural basis in the frontal lobes (e.g., Pribram, 1973; see historical reviews by Barkley, 2012 and Goldstein et al., 2014), it has been pointed out that EF also involves other brain areas and not all cognitive processes that employ the frontal lobes should necessarily be classified as EF (Barkley, 2012; see also Friedman & Robbins, 2022). Also, to define EF as what the frontal lobes or their prefrontal areas do and then state that what these brain areas do is EF results in a circular argument (Barkley, 2014), and thus a definition based solely on brain anatomy remains problematic.

To study which and how many separate cognitive constructs EF consists of, several authors have applied various statistical methods to empirical data to produce psychometric models of EF (Anderson, 2002; Grodzinsky & Diamond, 1992; Levin et al., 1996; Mariani & Barkley, 1997; Miyake et al., 2000; Shute & Huertas, 1990). A particularly influential psychometric model was presented by Miyake and colleagues (2000) based on adult data and later replicated with children by Lehto and others (2003). The model is often referred as the Miyake or Miyake-Friedman model and includes three basic EF components, namely shifting, updating, and inhibition (Miyake et al., 2000). However, other authors using different sets of EF measures have ended up with different models with a different number of latent variables and different descriptions (e.g., Anderson, 2002; Grodzinsky & Diamond, 1992; Levin et al., 1996; Mariani & Barkley, 1997; Shute & Huertas, 1990). Because of this, several authors have argued that these studies do little to shed light on the structure of EF, as the results depend on what tasks are considered EF measures to begin with (e.g., Andrewes, 2016; Barkley, 2012), and the differences between the results could simply reflect variation in task choice between the studies (Barkley, 2012). Furthermore, the psychometric models based on "cold" cognitive tasks have been criticized for neglecting emotional influences in goal-directed behavior (Barkley, 2012; see also Baggetta & Alexander, 2016), and their adequacy in providing satisfactory explanations for

empirical findings as well as specific, testable hypotheses has been questioned (e.g., Doebel, 2020; Perone et al., 2021; see also Burgess et al., 2006).

To conclude, there are numerous different definitions, theories, and models of EF that vary in their scope and origins (Baggetta & Alexander, 2016; Barkley, 2012; Goldstein et al., 2014; Jurado & Rosselli, 2007). However, all of them converge in that EF are seen as processes that are needed for effortful, goal-directed behavior. To fit the concept to a larger whole, EF can also be regarded as the cognitive top-down processes in a larger concept of self-regulation, which encompasses both top-down and bottom-up regulatory processes of action, emotion, and cognition during goal-directed behavior (Hofmann et al., 2012; Doebel, 2020; Nigg et al., 2017).

1.1.2 Prospective memory

PM, the memory for activities to be performed in the future (Einstein & McDaniel, 1990; Kliegel et al., 2008b), is another concept related to goal-directed behavior. These activities are intentions that cannot be completed immediately when they are formed but only several minutes, hours, days, or even weeks later (Rummel & Khavilashvili, 2022). The earliest PM studies tended to focus on remembering intentions in naturalistic everyday life settings, but later experimental paradigms have also been devised (Burgess et al., 2011). Akin to EF problems, PM failures can have a major impact on independent daily functioning of the individual, and they are very common complaints in clinical interviews (Einstein and McDaniel, 2005; McDaniel & Einstein, 2007). Thus, PM is important across the lifespan (Ballhausen et al., 2019). PM matures throughout childhood (Kvavilashvili et al., 2008, Mahy et al., 2014) and the development of EF and PM are intertwined (Zuber et al., 2019). Like EF, PM is also critical for gaining autonomy and independence from parents and caregivers (Cottini, 2023; Zuber et al., 2019).

When related to goal-directed behavior, PM can be seen as a form of goal-directed behavior where the aim is to perform the right action at the right time in the future. In *event-based PM* (EBPM), this is dictated by an external event (e.g., seeing a friend prompts you to return the money you borrowed earlier), whereas in *time-based PM* (TBPM), the given action should be performed at a specific time of the day or after a particular elapsed time (e.g., leave for school at 8.45 a.m.; Einstein & McDaniel, 1990). Some researchers (e.g., Kvavilashvili & Ellis, 1996) also make note of activity-based PM, where the intention needs to be executed after the completion of another activity (e.g., to take medication after breakfast). In real life, TBPM tasks are more common than EBPM tasks (Rummel & Khavilashvili, 2022).

As with EF, several theories have been advanced to define the cognitive processes involved in PM. The three main theories of PM are the multi-process theory, preparatory attention and memory processes theory, and prospective memory decision control theory (Rummel & Khavilashvili, 2022). All three theories are based on the task type first introduced by Einstein and McDaniel (1990), called the dual-task paradigm, which consists of an actual PM task and an ongoing task in which the actual PM task is embedded. The first of the theories, the multi-process theory of PM, states that both effortful cue monitoring (a top-down process) and spontaneous retrieval (a bottom-up process) underlie successful PM performance (McDaniel & Einstein, 2000). In turn, the preparatory attention theory states that individuals always need to engage in preparatory processing prior to PM target, which requires top-down working memory resources (Smith, 2003). Prospective memory decision control theory suggests that a self-imposed strategic slowing of ongoing task responding allows the PM responses (Strickland et al., 2018). One key difference between the latter two theories is that the prospective memory decision control theory suggests that the processing of the two tasks will only interfere with each other when individuals work beyond their cognitive capacity limits (Rummel & Khavilashvili, 2022). In their current forms, the three theories seem increasingly similar with regard to the types of processes involved in PM (Rummel & Khavilashvili, 2022). All three theories consider mainly EBPM and do not include hypotheses about any of the specific time perception and time management processes required in TBPM (Rummel & Kavhilashvili, 2022). Thus, as is the case with EF, a comprehensive theory of PM that would cover its different forms remains to be postulated.

Regarding neural correlates, research shows a consistent relation between activation in rostral prefrontal cortex and performance in PM tasks (Burgess et al., 2011). Thus, as with EF, frontal lobes seem to be particularly important for PM, even though more widespread brain networks, for example, in the parietal lobe and anterior cingulate cortex, are involved in both (Burgess et al., 2011).

1.2 Construct-driven assessment of goal-directed behavior

The development of assessment methods for goal-directed behavior and its underlying cognitive processes, such as EF and PM, has been dominated by the so-called construct-driven approach (Burgess et al., 2006; see also Dawson & Marcotte, 2017; Parsons et al., 2017). In this approach, measures are developed and their results evaluated based on the latent cognitive constructs that they are hypothesized to tap (Burgess et al., 2006; Dawson & Marcotte, 2017; Parsons et al., 2017). As these hypothetical constructs cannot be directly observed, their existence is inferred from research findings such as correlations between different

measures (Burgess et al., 2006). These measures can be divided into performance-based measures and ratings, which are covered next.

1.2.1 Performance-based measures

As regards to performance-based measures, experimental psychology has heavily relied on simplified paradigms that comprise impoverished stimuli and limited behavioral responses and supposedly measure some given cognitive construct (Hatfield, 2002; see also Dawson & Marcotte, 2017). In test measures utilized in clinical neuropsychology, the patient is typically given a single explicit problem at a time in short trials while task initiation is strongly prompted by the examiner (Shallice & Burgess, 1991; see also Dawson & Marcotte, 2017). These tests and measures have the advantage of being explicit, objective and well-controlled, which helps to minimize any unwanted effects of confounding variables (Toplak et al., 2013; see also Hatfield, 2002). On the downside, these typical performance-based measures lack the abundance of contextual information that triggers affordances, prospective goals, and cognitive schemas characterizing everyday life (see Wilson, 2002). Depending on the features of the targeted cognitive process or construct, this may pose problems for the assessment in varying degrees. For example, the performance-based measures are subject to the task-impurity problem, which means that since any target process/construct needs to be measured within some specific task context, all scores necessarily include systematic variance attributable to other cognitive factors (e.g., Miyake & Friedman, 2012; Miyake et al., 2000). It has thus been questioned if these kinds of paradigms are sufficient for studying something as complex as human goal-directed behavior as it occurs in everyday situations (e.g., Bronfenbrenner, 1977; Gibson, 1970; Kingstone et al., 2008; Neisser, 1976).

Performance-based measures of EF are tasks that have been considered to measure EF either on theoretical grounds or after being found to be sensitive to cognitive impairments caused by prefrontal lesions (Burgess et al., 2006; Pennington & Ozonoff, 1996). Thus, they originate from varied conceptual and experimental frameworks (Burgess et al., 2006). Some of the most widely used EF tests include variants of the Wisconsin Card Sorting Test (Grant & Berg, 1948), the Stroop task (Stroop, 1935), the Trail Making Test (Armitage, 1946), and the Tower of London (Shallice, 1982). Many neuropsychological test batteries, such as the NEPSY-II (Korkman et al., 2007), and the Delis-Kaplan Executive Function System (Delis et al., 2001), either include implementations of these EF paradigms or consist solely of them.

Criticizing the performance-based measures of EF, Barkley (2012) argues that *"a myopic emphasis on short-term (minutes) cold cognitive psychometric tasks preferred by most contemporary neuropsychological studies has left a gapping*

chasm between the constructs sampled by these tests and the executive deficits evident in patients in their everyday life". Also others have questioned the utility of these measures in predicting how well individuals will do with EF in their daily life activities (e.g., Burgess et al., 2006; Chaytor and Schmitter-Edgecombe, 2003; see also Chan et al., 2008). Because of the simplified nature of many of the EF tasks and low correlations between them and EF ratings (e.g., Barkley, 2012; Toplak et al., 2013), it remains unclear how the results of EF tasks are linked to everyday life and what implications they may have to the real-life functioning of an individual (Burgess et al., 2006), thus resulting in limited utility in the evaluation of daily adaptive functioning (Barkley, 2012).

Features of a typical performance-based PM measure has been described by Burgess, Scott and Frith (2002; see also Burgess et al., 2011; Rummel & Kvavilashvili, 2022) and can be summarized as follows: 1) there is an intention that cannot be performed immediately, but needs to be performed in a particular circumstance, and the delay between creating the intention and the appropriate time to act is filled with an ongoing task; 2) the performance of the ongoing task prevents continuous rehearsal of the intention over the delay; 3) the intention cue (i.e., the particular circumstance) does not interfere with performance of the foreground task; 4) in most situations, no immediate feedback is given to participants regarding PM performance. Probably the most influential PM task paradigm has been the dual-task paradigm that was introduced by Einstein and McDaniel (1990). In a typical dual task paradigm, the ongoing task is simple and appropriate moment to perform the PM task is signaled by a certain event (EBPM) or a given time (TBPM; see section 1.1.2). After the study of Einstein and McDaniel (1990), numerous different dual-task variants have been introduced, and the main theories of PM have also been influenced by the dual task paradigm (Rummel & Khavilashvili, 2022). Some PM tests intended for clinical use other tasks, which yield clinically relevant information in their own right, as ongoing task, which can save valuable evaluation time (See Thöne-Otto & Walther, 2008).

As a response to recognizing that everyday life PM is employed under complex environmental factors, also complex PM tasks have been devised (e.g., Kliegel et al., 2000; Kliegel et al., 2006). Interestingly, some of these complex PM tasks, such as the MET and Six Element Test (Shallice & Burgess, 1991), were first developed to study EF deficits but have successfully been adopted to study PM (Burgess et al., 2000; Kliegel et al., 2000; Kliegel et al., 2004; Kliegel et al., 2006). Several other studies also provide converging evidence about the close relationship between EF and PM performance measures (e.g., Groot et al., 2002; Kliegel et al., 2003; McDaniel et al., 1999). Thus, even though performance-based measures can be designed in a construct-driven manner, that is, targeting a specific area of cognition, it can remain ambiguous which cognitive constructs are actually being quantified.

1.2.2 Ratings

Ratings like questionnaires provide another way for measuring goal-directed behavior or hypothesized cognitive processes. With adults, these are usually self-reports or are filled out by the spouse/family member, while with children, teacher reports are also common. Ratings are able cover areas of goal-directed behavior that can be difficult to assess with traditional performance-based measures, for example the ability to pursue goals that span over longer time periods, such as weeks or months. Toplak and colleagues (2013) suggest that EF questionnaires are indicative of success in goal pursuit, whereas performance-based measures capture the efficiency of cognitive abilities. This distinction shares resemblance to a suggestion made by Barkley (2012), who states that performance-based EF tests, despite with their shortcomings, can be used to assess the instrumental level of EF. In contrast the methodological, tactical, and strategic levels of his extended phenotype model of EF are more successfully captured by ratings.

According to Toplak and colleagues (2013), *the Behavior Rating Inventory of Executive Function* (BRIEF; Gioia et al., 2000) has been the most common EF questionnaire. Originally, the BRIEF was developed to provide an ecologically valid alternative for EF tests, while also aiming to quantify EF aspects that may not be covered by these performance-based measures (Gioia et al., 2000; Roth et al., 2013). In some other EF questionnaires the design can be seen to more closely assess the same cognitive constructs as the EF tasks (Toplak et al., 2013). For example, the Childhood Executive Functioning Inventory has only scales for inhibition and working memory, which are measured by many performance-based tests (Thorell et al., 2010; Thorell & Nyberg, 2008). Other common questionnaires probing EF include the Attention and Executive Function Rating Inventory (Klenberg et al., 2010), Behavioural Assessment of the Dysexecutive Syndrome (Norris & Tate, 2000), and 5–15 questionnaire (Kadesjö et al., 2004), which also cover other cognitive processes and behaviors.

Also PM can be evaluated with ratings, such as the Prospective Retrospective Memory Questionnaire (Crawford et al., 2003) and Metacognitive Prospective Memory Inventory (Rummel et al., 2019). Also diaries, where the respondents on several occasions make notes of their PM performance, are commonly used in adults (e.g., Haas et al., 2020; Jylkkä et al., 2023a; Jylkkä et al., 2023b). It should be noted that some questionnaires not originally designed to measure PM have items that may clearly be regarded as measuring PM. For example, the BRIEF includes questions about failures to fulfill previously planned intentions (Gioia et al., 2000), and *the ADHD Rating Scale-IV* (ADHD-RS; DuPaul, 1998) has items about being forgetful in daily activities.

Like performance-based measures, ratings have their own shortcomings. Most notably, the evaluation is subjective and may be biased because of the personal

views of the assessor or because the assessor may not know the assessee well enough to make informed evaluations. This can be the case, for example, when a teacher is asked to fill out a questionnaire about a pupil who just recently has joined the class. Also, the assessment of the severity of the possible symptoms could be particularly difficult. This is because questionnaires tend to have vague rating scales, such as, ”is never a problem”/”is sometimes a problem”/”is often a problem”, with no accurate way of defining at what point, for example, “sometimes” becomes “often”. The questionnaires often ask the respondent to make evaluations about behaviors that occur over a long period of time. For instance, in the BRIEF and ADHD-RS this period is last six months (DuPaul, 1998; Gioia et al., 2000). This may be an advantage as it increases the possibility that the respondent has witnessed the behavior of the assessee in an adequate number of situations, but remembering incidents that happened almost half a year ago or making good generalizations from longer time spans can be difficult. The behavior can also fluctuate very much, either as a function of time or depending on the context. In some cases a suitable person to fill out the questionnaire may simply not exist. Finally, the way the questionnaire or its individual items are designed or administered can produce bias in multiple ways (Choi & Pak, 2005).

One influential challenge in the construct-driven approach is shared by both performance-based and rating measures. As mentioned in section 1.1.1, there is no general of agreement on EF and what specific cognitive constructs it consists of. Therefore, the selection of a constellation of construct-driven tasks for a comprehensive assessment of EF remains challenging. Together with the task-impurity problem (e.g., Miyake et al., 2000; Miyake & Friedman, 2012), this makes the interpretation of construct-driven measures difficult, especially in clinical settings (see Burgess et al., 2006).

1.3 Ecological validity

As mentioned in section 1.2, the construct-driven tests and measures of goal-directed behavior have been criticized for not being representative of the contexts and challenges of real life and providing little predictive value for everyday executive and goal-directed behaviors (Barkley, 2012; Burgess et al., 2006; see also Chaytor and Schmitter-Edgecombe, 2003; Parsons, 2015). Thus, these tests are considered to have low ecological validity. The origins of the term ecological validity can be traced back to Brunswik (1943), but the current use of ecological validity has drifted away from the original context (see, e.g., Bronfenbrenner, 1977; Chaytor & Schmitter-Edgecombe, 2003; Franzen and Wilhelm, 1996; Kvavilashvili and Ellis, 2004; Spooner & Pachana, 2006).

Franzen and Wilhelm (1996) recognize two different types of ecological validity, namely *verisimilitude* and *veridicality*. They define verisimilitude as how

accurately the test properties resemble the corresponding situations in the real world and veridicality as the extent a test reflects or predicts the skills or performances in everyday life (see also Chaytor & Schmitter-Edgecombe, 2003; Spooner & Pachana, 2006). This definition bears resemblance to the one used by Kvavilashvili and Ellis (2004), who define ecological validity as consisting of representativeness (i.e., verisimilitude) and generalizability (i.e., veridicality). The terms verisimilitude and veridicality have also been applied to describe two ways for improving ecological validity (Chaytor & Schmitter-Edgecombe, 2003; Spooner & Pachana, 2006). In the verisimilitude approach, new tests are designed to simulate critical everyday cognitive tasks (Chaytor & Schmitter-Edgecombe, 2003; Spooner & Pachana, 2006). The veridicality approach, in turn, involves using existing tests and optimizing their measures to be optimally related to real-world functioning (Chaytor & Schmitter-Edgecombe, 2003; Spooner & Pachana, 2006).

To attain higher ecological validity, goal-directed behavior has been evaluated with tasks that take place in actual real-life locations, such as a shopping street, instead of laboratory or hospital rooms (e.g., Garden et al., 2001; Shallice and Burgess, 1991). These tasks have very high verisimilitude, as they are conducted in actual real-life environments (Spooner and Pachana, 2006). However, experiments conducted in real-life settings come with several obvious drawbacks (Parsons, 2015). For example, they can be difficult to standardize or replicate, because unpredictable events and behavior of other people may affect the results, as the researchers have only limited control over the environment. The precise measurement of behavior can also be challenging, even though new methodologies, such as ecological momentary assessment, that is, repeated sampling of subjects' behaviors in natural environments, have produced advancements in this matter (see, Shiffman et al., 2008). Experiments in real-life locations may not always be suitable for patients with mobility, behavioral, or psychiatric problems (Knight et al., 2002). Therefore, despite the benefits, using real-life environments may in some cases result in low reliability and participant safety can be compromised (Logie et al., 2011). Using real-life environments can also be time-consuming and impractical.

A less laborious approach is to include real-life features to laboratory tasks. In the Virtual Week task, participants are asked to perform everyday tasks in a board game (Rendell and Craik, 2000). However, whereas tasks performed in real-life locations have been shown to better predict behavioral everyday difficulties than the traditional EF tests (Alderman et al., 2003), the same remains to be shown for laboratory tasks depicting everyday situations.

Yet another attempt to improve the ecological validity of EF assessments has been to complement the results of EF tests with questionnaires regarding everyday compensatory strategies and environmental cognitive demands (Chaytor

et al., 2006). Miller and Barr (2017) suggest that methods quantifying real-time behavior in the real world alongside information about the local context could substantially improve the ecological validity of cognitive assessments. Furthermore, they criticize the field of neuropsychology of being reliant on outdated data collection methods that do not, unlike those employed in most other fields of medicine, take advantage of rapid technological innovations. Applying new technological advancements could result in better treatment recommendations and thus improved patient outcomes (Miller & Barr, 2017).

The issue of ecological validity is very relevant to both scientific research and clinical work. Virtually all psychiatric and neurological disorders and conditions have everyday life implications, and a core goal in neuropsychological rehabilitation is to enhance quality of life (Dawson & Marcotte, 2017). Traditional cognitive tasks often require discrete responses to singular events in carefully controlled environments, whereas performance in the real world involves parallel streams of tasks, often in disordered environments (Dawson & Marcotte, 2017; see also Hatfield, 2002; Wilson, 2002). This discrepancy between traditional tasks and real-life functions may significantly limit the ecological validity of these measures (Barkley, 2012; Dawson & Marcotte, 2017; Levine et al., 2000). All in all, new assessment methods with higher ecological validity or improved versions of the old ones are needed, especially for studying goal-directed behavior and related concepts like EF and PM.

1.4 Function-led assessment of goal-directed behavior

To overcome the challenges in the construct-driven assessment methods of goal-directed behavior, alternative approaches have been proposed (e.g., Burgess et al., 2006; Kingstone et al., 2008; Parsons et al., 2017). These approaches converge in that for attaining higher ecological validity, task design should be based on real-world behaviors and functions rather than hypothetical cognitive constructs.

In the function-led task design approach, the rationale is to emulate real-world functions and the environments in which these functions are used to reach increased ecological validity (Burgess et al., 2006; Parsons et al., 2017). This approach can be taken to produce measures to be performed either in the laboratory, such as the Six Element Test (Shallice & Burgess, 1991; see also Burgess et al., 1996; Kliegel et al., 2000), or in other environments, like *the Multiple Errands Test* (MET) and its variants (Shallice & Burgess, 1990; see also Rotenberg et al., 2020). Such tests should be validated against measures of ability in the real world, such as observer or self-ratings (Burgess et al., 2006). Burgess and colleagues (2006) argue that function-led tests may better be suited to clinical use than construct-driven tests because of the transparency afforded by increased representativeness (i.e., verisimilitude) and generalizability (i.e., veridicality) for

real-life functioning, while being psychometrically as robust as construct-driven tests. As regards to generalizability, for example, a hospital version of the MET has been found to correlate with self-report measures of everyday ability (Dawson et al., 2005), objective assessment of daily living skills (Dawson et al., 2005), and caretakers' ratings of everyday life problems (Knight et al., 2002). In their meta-analysis, Rotenberg and collaborators (2020) found strong evidence for some psychometric properties of the MET, such as validity for discriminating certain clinical groups (e.g., dementia and acquired brain injury) from control groups. However, they also found that more research is needed on certain other psychometric aspects, such as test-retest stability and internal consistency. Burgess and collaborators (2006) underline that the argument should not be taken as suggesting the sole use of tests that are carried in the real-world settings, but that these new tasks are complementary as they could cover aspects of real-life situations that are not accounted for by the current construct-driven tests (see also Parsons, 2015). An obvious challenge for the MET and other tasks performed in real-life environments is how to maintain sufficient experimental control. This is needed for participant safety and to ensure reliability, which remains to be insufficiently studied for the MET (Rotenberg et al., 2020).

The function-led approach has also been utilized with children (e.g., Chevignard et al., 2010; Rocke et al., 2008). Chevignard and colleagues (2010) devised the Children's Cooking Task, where the participants prepare two simple recipes in a hospital kitchen. In this task, outcome measures comprise the number of errors and overall qualitative performance analysis, in which performance time, whether the child was able to prepare both recipes or not, and evidence for any dangerous behaviors, was considered. The authors found high internal consistency and test-retest stability for these measures. The Children's Cooking Task has been found to discriminate children with traumatic brain injury (Chevignard et al., 2010; Finnanger et al., 2022; Fogel et al., 2020), developmental dyspraxia (Toussaint-Tohrin et al., 2013), or neurodevelopmental disorders (Fogel et al., 2020) from typically developing controls. As the task is performed in a dedicated room, there is more control to the environment that with tasks that take place in larger environments like pedestrian areas (cf. Shallice & Burgess, 1990). However, such a specialized room may not be readily available and could need to be standardized to some extent.

Advocating for an approach coined as cognitive ethology, Kingstone and colleagues (2008) argue that laboratory research of human cognition is founded on two principles they view as unrealistic, namely the invariance assumption and the attempt to maximally control the test situation. The invariance assumption means that human cognition is underpinned by processes that are constant across situations, and by exercising control over the test situation, the researcher attempts to ensure that any change can be attributed to the variable that is being

manipulated. The authors proceed to point out that even very minor changes to a typical restricted laboratory paradigm can compromise the replicability of an effect, a fact that raises serious doubts as to whether the findings obtained using simplified laboratory tasks have utility in understanding real-world behavior. As a remedy, they suggest the cognitive ethology approach, where the aim is to first directly observe how people behave in their real-world environments, and once this complex problem space is sufficiently characterized, move into the laboratory to test hypotheses that are generated based on the real-world observations. This means moving away from making causal claims about fundamental cognitive processes towards observing and describing behavior as it occurs in real life (Kingstone et al., 2005). Overall, approaches like function-led task design and cognitive ethology can be seen as attempts to provide tools for understanding human behavior in its natural environments.

1.5 ADHD and its behavioral indications

The syndrome at the focus of this Thesis, *Attention-deficit/hyperactivity disorder* (ADHD), is one of the most common neurodevelopmental disorders in childhood (Bitsko et al., 2022) with prevalence of 5 to 11 percent in persons under 18 years of age (Francés et al., 2022). ADHD is characterized by symptoms that fall on the three dimensions of inattention, hyperactivity, and impulsivity (World Health Organization, 2016). The diagnostic criteria for ADHD according to the Tenth Edition of the Internal Classification of Diseases (ICD-10; World Health Organization, 2016) are presented in Table 1.

Depending on which of the three symptom dimensions are predominant in individual's behavior, three ADHD subtypes can be recognized, namely predominantly inattentive, predominantly hyperactive/impulsive, and combined subtype. ADHD is linked with several adverse outcomes, such as impairments in quality of life, emotional and social challenges, and educational underachievement (Agnew-Blais et al., 2018; Faraone et al., 2021). Comorbid diagnoses are common (Faraone et al., 2021; Francés et al., 2022). Furthermore, individuals with ADHD often display problems with EF (see, e.g., Barkley, 2012; Barkley & Murphy, 2011) and PM (e.g., Kliegel et al., 2008a; Talbot et al., 2018).

As ADHD often is a lifelong disorder and interventions can reduce its adverse consequences, timely diagnostics with accurate and objective triaging of the core symptoms is vital (Sayal et al., 2018; Sonuga-Barke & Halperin, 2010). As of today, diagnosing ADHD relies mainly on interviews and questionnaires that are prone to subjective bias (Gualtieri and Johnson, 2005), while the predictive value and clinical utility of objective experimental test methods (Hall et al., 2016; Nichols and Waschbusch, 2004) and possible biomarkers (Mehta et al., 2020; Thome et al., 2012) is being investigated. Next, the effects of ADHD on task-

performance measures and motion- and eye tracking-based assessment of ADHD symptoms are discussed.

Table 1 The diagnostic criteria for ADHD according to the Internal Classification of Diseases, Tenth edition.

G1. Inattention. At least six of the following symptoms of inattention have persisted for at least six months, to a degree that is maladaptive and inconsistent with the developmental level of the child: (1) Often fails to give close attention to details, or makes careless errors in schoolwork, work, or other activities. (2) Often fails to sustain attention in tasks or play activities. (3) Often appears not to listen to what is being said to him or her. (4) Often fails to follow through on instructions or to finish schoolwork, chores, or duties in the workplace (not because of oppositional behaviour or failure to understand instructions). (5) Is often impaired in organizing tasks and activities. (6) Often avoids or strongly dislikes tasks, such as homework, that require sustained mental effort. (7) Often loses things necessary for certains tasks or activities, such as school assignments, pencils, books, toys, or tools. (8) Is often easily distracted by external stimuli. (9) Is often forgetful in the course of daily activities.
G2. Hyperactivity. At least three of the following symptoms of hyperactivity have persisted for at least six months, to a degree that is maladaptive and inconsistent with the development level of the child: (1) Often fidgets with hands or feet or squirms on seat. (2) Leaves seat in classroom or in other situations in which remaining seated is expected. (3) Often runs about or climbs excessively in situations in which it is inappropriate (in adolescents or adults, only feelings of restlessness may be present). (4) Is often unduly noise in playing or has difficulty in engaging quitely in leisure activities. (5) Exhibits a persistent pattern of excessive motor activity that is not substantially modified by social context or demands.
G3. Impulsivity. At least one of the following symptoms of impulsivity has persisted for at least six months, to a degree that is maladaptive and inconsistent with the developmental level of the child: (1) Often blurts out answers before questions have been completed. (2) Often fails to wait in lines or await turns in games or group situations. (3) Often interrupts or intrudes on others (e.g., butts into others' conversations or games). (4) Often talks excessively without appropriate response to social constraints.
G4. Onset of the disorder is no later than the age of seven years.
G5. Pervasiveness. The criteria should be met for more than a single situation, e.g., the combination of inattention and hyperactivity should be present both at home and at school, or at both school and another setting where children are observed, such as a clinic. (Evidence for cross-situationality will ordinarily require information from more than one source; parental reports about classroom behavior, for instance, are unlikely to be sufficient.)
G6. The symptoms in G1–G3 cause clinically significant distress or impairment in social, academic, or occupational functioning.
G7. The disorder does not meet the criteria for pervasive developmental disorders (F84.–), manic episode (F30.–), depressive episode (F32.–), or anxiety disorders (F41.–).

1.5.1 Task-performance measures

When studied using conventional, construct-driven paradigms, ADHD is associated with impairments in multiple cognitive domains, such as working memory, response inhibition, decision making, and attention (Pievsky and McGrath, 2018). This is hardly surprising, given that ADHD symptom dimensions include inattention and impulsivity, which attention and inhibition tasks should logically tap. Still, cognitive impairments in ADHD do not necessarily fall into categorical cognitive domains as defined by conventional test paradigms (Willcutt et al., 2005), but instead manifest themselves in heterogeneous and idiosyncratic patterns (Luo et al., 2019).

Perhaps the test paradigm with most consistent findings in differentiating children with versus without ADHD is *the Continuous Performance Task* (CPT; Albrecht et al., 2015; Ogundele et al., 2011; see also Berger et al., 2017). The CPT was originally developed to study brain damage (Rosvold et al., 1956). Later, several CPT variants have been developed that use either visual and/or auditory stimuli (see, e.g., Gualtieri & Johnson, 2005). Even though originally used for other purposes, the CPT has been considered useful as an adjunct to clinical diagnosis of ADHD and in the neuropsychological assessment of individuals with ADHD (Fuermaier et al., 2019; Gualtieri & Johson, 2005).

Accumulating brain imaging findings support the hypothesis that ADHD is not primarily a problem of cognitive resources but rather a problem of excessive fluctuation of how these resources are successfully used (e.g., Sonuga-Barke and Castellanos, 2007). If this is the case, variability measures that account for possible fluctuation of performance could better capture the cognitive manifestations of ADHD than sum scores, such as the mean reaction time or the total number of correct responses. This hypothesis is supported by task-performance findings showing that in various neuropsychological tasks, ADHD is more related to increased reaction time variability than reaction time itself (Kofler et al., 2013; Tamm et al., 2012; see also Pievsky and McGrath, 2018).

Regarding the assessment of ADHD symptoms, an important drawback in these conventional, construct-driven paradigms relates to their highly structured nature and the assumption that maximal performance in a simplified task with limited behavioral responses is an informative predictor of how these symptoms manifest in the complex and varied situations of everyday life (Barkley and Murphy, 2010; Parsons et al., 2017). Monotonous and highly structured tasks where the participants are forced to constantly work at their capacity limits is not akin to most situations where goal-directed behavior occurs in real life. Rather, real-life goal-oriented behavior is composed of dynamically changing cascades of daily actions and affected by the interplay between the individual and the environment (Ackerman, 1994; Toplak et al., 2013). Also, real-life environments are abundant with contextual information that trigger, support, and sometimes

distract goal-directed behavior (e.g., Marsh et al., 2008), which is lacking in simplified tasks that use restricted stimuli. It is therefore not surprising that measures and behavioral observations in such tasks have limited predictive value to the real-life settings where the children with ADHD live and where their symptoms emerge (Barkley, 1991; Barkley & Murphy, 2010, 2011; Grodzinsky & Barkley, 1999; Hall et al., 2016). To summarize, there is a need for measures that would quantify ADHD symptoms in rich, open-ended, and dynamic environments akin to real life (e.g., Kingstone et al., 2008).

1.5.2 Motion-based measures

Of ADHD diagnostic criteria presented in Table 1, those falling into the dimension of hyperactivity are perhaps most straightforward to measure, as most of them are reflected in gross motor activity. Previously, activity levels of persons with ADHD have been quantified using various sensor technologies worn in different body parts like wrist, waist, and ankle (see meta-analysis by De Crescenzo et al., 2016). Naturalistic motion tracking studies have the challenge of distinguishing abnormal or non-adaptive motion patterns from typical overall activity levels and variations caused by contextual factors. There are also studies that measure participant movement during cognitive tasks like the CPT (e.g., Nolin et al., 2012; Mangalmurti et al., 2020; Teicher et al., 1996; see also Parsons et al., 2019). Tasks used in these studies often call for sitting in a fixed position and require only minimal physical movement, such as pressing a single button. This means that despite their indisputable achievements, these tasks may not be optimally representative of real-life physical activity. It has been suggested that hyperactivity would be most clearly observed in cognitive tasks where the level of stimulation is low (Kofler et al., 2016). While there is no clear evidence to dispute this, paradigms that call for more naturalistic and varied movements could offer better representativeness (i.e., verisimilitude) of everyday situations.

1.5.3 Eye movement behaviors

The attentional difficulties in ADHD are reflected in different eye movement behavior aspects, such as saccades and fixations. Saccades are ballistic eye movements that redirect gaze, and between them, during fixations, gaze is held almost stationary by slow stabilizing movements (Liversedge et al., 2011). Reflecting visual attention control and depth of the visual information processing, studies with restricted tasks have reported group differences between children with ADHD and their typically developing peers (Caldani et al., 2019; Fried et al., 2014; Karatekin, 2007; Levantini et al., 2020; Mohammadhasani et al., 2020). Based on these studies, ADHD is associated with deficits in goal-directed

attention, such as difficulties in maintaining fixation on a target stimulus (Karatekin, 2007), less accurate saccades to target location (Karatekin, 2007), and reduced ability to suppress saccades (Caldani et al., 2019; Fried et al., 2014; Levantini et al., 2020). Children with ADHD also exhibit shorter saccade latencies, lower saccade peak velocities, and less accurate saccade landing (Bucci et al., 2017; Castellanos et al., 2000; Karatekin, 2007). Regarding stimulus-driven attention, that is, engagement of attention via some external factor in the environment (Corbetta & Shulman, 2002), some studies have indicated that children with ADHD show a bias toward salient locations (e.g., Tseng et al., 2013). While these studies with restricted stimuli shed light on attention-related characteristics of ADHD, it remains unresolved if the results concerning the aberrant eye movement behaviors in ADHD generalize to real-life situations. To identify the attention-related mechanisms of ADHD symptoms in real life, eye movements could be studied in more naturalistic tasks where the participants are able to move their head, navigate freely, and interact with objects (Parsons et al., 2019). This could be accomplished with head-mounted displays (see section 1.6), but previous ADHD studies with this technology have either not used naturalistic tasks or have focused on gaze and scan path analyses rather than on an in-depth analysis of saccades and fixations (Mangalmurti et al., 2020; Sitzmann et al., 2018; Stokes et al., 2022).

In addition to being related to total amount of eye movement in a given time, ADHD manifests itself as increased fluctuations in attention over time (e.g., Cheung et al., 2017). In visual attention research, a distinction between ambient (long saccades and short fixations scattered over wider area) and focal (shorter saccades and longer fixation within a smaller area) mode has been made (Eisenberg & Zacks, 2016; Holm et al., 2021). With static stimuli, ambient mode usually changes to focal mode over time, but with dynamic stimuli the participants start to alternate between the two modes (Eisenberg & Zacks, 2016; Velichkovsky et al., 2002). Reaction time variability is one of the most promising behavioral markers of ADHD (Kofler et al., 2013; Tamm et al., 2012), but it remains unclear if this fluctuation also happens in eye movement behavior such as mode switching.

Visual saliency indicates how much a particular visual area stands out from its environment in color, orientation, intensity, or other low level visual characteristics (Itti & Koch, 2001; Parkhurst et al., 2002). In absence of an active task, saliency is considered a main factor to guide human attention (Itti & Koch, 2001; Judd et al., 2012; Parkhurst et al., 2002). During task performance, salient but irrelevant stimuli can interfere with task performance (Theeuwes & Belopolsky, 2012). Based on previous research, individuals with ADHD tend to orient their gaze to highly salient stimuli more than their typically developing peers (Tseng et al., 2013; Shalev et al., 2010), which can manifest as higher rate of saccades towards location of salient stimuli and longer fixation times on them.

This can also be observed, for instance, as higher error rate in an anti-saccade task (Bucci et al., 2017; Fernandez-Ruiz et al., 2020; Mostofsky et al., 2001). Moreover, neuroimaging studies suggest abnormal function of the salience network in the brain of individuals with ADHD (Cortese et al., 2012). Stimulus-driven (bottom-up, saliency-related) shift of attention and voluntary goal-directed (top-down) attention regulation are two major attentional mechanisms that guide selective visual attention (Parkhurst et al., 2002). Therefore, the number and duration of fixations to salient locations can indicate distraction during active task performance (Geng & DiQuattro, 2010; Born et al., 2011), which could potentially be used to quantify attentional lapses in children with ADHD (Stokes et al., 2022).

1.6 Virtual reality

Virtual reality (VR) can be defined as *"an advanced computer interface that allows the user to interact and become immersed within computer-generated simulated environments"* (Rizzo et al., 1997). The first papers describing potential computer-generated environments akin to immersive VR were published in the 1960s (e.g., Sutherland, 1965; see also Cipresso et al., 2018; Slater and Sanchez-Vives, 2016). In the 1980s, the term VR was used by Jaron Lanier, who has been credited for applying the term in the way it is known today (Slater and Sanchez-Vives, 2016). In the 1990s, VR hardware was still very modest compared to the standards of today, but there already were commercial products in several areas, such as vehicle simulation, medicine, probe microscopy, and architectural design (Brooks, 1999). Researchers also acknowledged the potential of VR in clinical psychology (Riva, 1997) and cognitive rehabilitation (Rizzo et al., 1997). Moreover, the development of some landmark task paradigms in the field of VR-based psychological research, such as the Virtual Classroom, began (Rizzo et al., 2009).

During this millennium, the number of scientific publications encompassing VR has increased rapidly (Krohn et al., 2020). Already in 2016, there already was a notable number of 186,000 scientific VR papers to be found in Google Scholar (Slater and Sanchez-Vives, 2016). This scientific development has been accompanied by increased attraction from investors and the general public, especially after Facebook (Meta from 2021 onwards) acquired Oculus company for two billion dollars in 2014 (Luckerson, 2014). After this, many other major companies, such as HTC, Sony, Google, and Samsung, have made formidable investments in VR (Castelvecchi, 2016), which had led to rapid improvements in the low-cost VR technology. Up to this day, VR has been applied in a wide range of fields, such as gaming, psychotherapy, education, social skills training, simulations of surgical procedures, military training, and architectural design (Cipresso et al., 2018; Slater & Sanchez-Vives, 2016). The potential of VR for widespread clinical use has also been acknowledged (Rizzo & Koenig, 2017).

Next, the features of immersive and non-immersive VR platforms are compared, and the studies comparing these two platforms are reviewed. As the focus of this Thesis is on immersive VR, this is followed by a review of psychological studies with tasks designed either construct-driven or function-led manner and implemented using immersive VR technology.

1.6.1 Immersive and non-immersive VR

VR can be implemented with various technical solutions that differ in their immersiveness, such as traditional *flat screen displays* (FSDs) with traditional interaction devices (e.g., keyboards, gamepads, and mice) or *head-mounted displays* (HMDs; Cipresso et al., 2018; Di Natale et al., 2020) with related hand controllers. An immersive VR system is one that allows the participant to perceive the environment and interact with it through natural sensorimotor contingencies (Slater & Sanchez-Vives, 2016). Alternatively, an immersive VR system can be defined as a system that blurs the lines between the physical and virtual worlds (Suh & Prophet, 2018). High immersiveness calls for effective sensory substitution, which depends on factors such as a wide *field-of-view* (FOV), using head tracking for changing the FOV, a short latency from head move to display change, a high display resolution, and stereoscopic vision and sound (Slater & Sanchez-Vives, 2016). As a broad categorization, the HMD-based VR systems and related position-tracking controllers and camera-based hand tracking can be regarded as immersive VR, and the systems using FSDs and more traditional controllers as non-immersive VR (e.g., Di Natale et al., 2020; Suh & Prophet, 2018). It should be noted that there are also semi-immersive VR technologies, such as the cave automatic virtual environment, where walls, ceiling and floor are covered with projected images (see, e.g., Di Natale et al., 2020). The immersiveness of VR systems should therefore be regarded as a continuum, not a dichotomy. In this Thesis, only FSD- and HMD-based VR systems were used and only these technologies are covered in greater detail.

The sense of presence, a subjective correlate of immersion, is the illusion of "being there" in the VR environment while being aware that this is not actually true (Slater & Sanchez-Vives, 2016). The sense of presence can be considered a key aspect of VR experience and its ecological validity in terms of verisimilitude, as it can be argued that only when the participants are experiencing a strong sense of presence, they will perform the task as they would do them in real life (Pan & Hamilton, 2018; Slater, 2018) and show similar reactions to the environment as in real world (Kothgassner & Felnhofer, 2020).

Systems that use HMDs have several advantages over those employing traditional FSDs and the related peripherals, as they can more closely emulate real-life sensorimotor experiences than FSDs and therefore better match the

criteria for an immersive VR system. For example, turning and rotating the head with an HMD alters the view in the VR world in parallel with the actual physical movement, which cannot be achieved with typical FSDs. The hand controllers of the current HMD hardware, such as Pico Neo 3 or Oculus (Meta) Quest 2, track their position and rotation, hence moving and turning the physical controller leads to similar movements and rotations in the controller object seen in the virtual space. The FOV in the current HMDs is also considerably larger than in the traditional FSDs. As an example, Pico Neo 3 Pro HMD has a horizontal FOV of 98 degrees, while a 27" FSD with an aspect ratio of 16:9 produces a horizontal FOV of 41 degrees when viewed from 80 cm. Systems with HMDs produce a stereoscopic visual experience (Parsons, 2015) and usually block the view of the surrounding physical environment completely, which can further increase immersiveness (see Slater, 2018). These advantages in HMDs can lead to higher sense of presence (Caroux, 2023; Chang et al., 2020; Li et al., 2020; Makransky et al., 2019; Pallavicini et al., 2019; Pallavicini & Pepe, 2019; Pallavicini et al., 2018; Tan et al., 2015; Yao & Kim, 2019) and behavioral consequences, such as greater physical effort (e.g., Yao & Kim, 2019).

There are also potential drawbacks with HMDs compared to FSDs. The earlier HMD models released before 2010 were sometimes reported to cause cybersickness symptoms (Bohil et al., 2011) which include nausea, disorientation, and oculomotor symptoms (Kourtesis et al., 2023). However, these adverse effects are markedly smaller or absent with HMD models released in 2016 or later (onwards from Oculus Rift and HTC Vive; Kourtesis et al., 2019a; Weech et al., 2019). Eliminating cybersickness is essential not only for the comfort of the participant but also for ensuring ecological validity, as cybersickness and the sense of presence are negatively associated (Weech et al., 2019). Recent studies show that cybersickness is related to display lag in virtual and physical head pose (Palmisano et al., 2022) and that it can be countered with dynamic FOV restriction (Teixeira & Palmisano, 2021). Despite technological improvements, cybersickness could arise in situations where there is a conflict between perceived and physical movements (Bohil et al., 2011; Palmisano et al., 2020; Palmisano et al., 2022), and some individuals may be particularly sensitive to it (Parsons et al., 2017). Because of potential cybersickness symptoms, FSDs could still be the preferred choice in some situations, such as in wheelchair training (Alapakkam Govindarajan et al., 2022) or in race driving simulations (Walch et al., 2017). Accordingly, Parsons and others (2019) note that due to the sensory issues of many individuals with autism spectrum disorder, less invasive methods of presenting virtual environments than HMDs should also be studied. As typical laptop and desktop computers with FSDs and related peripherals can be used as non-immersive VR systems, the required technology is widely available, and even less technically oriented users have been accustomed to use the related user

interfaces and operating systems. The HMD-based VR systems are still less widely available than FSDs, and the user could need additional training to use them, if one does have no prior experience using the hardware and related software. Overall, the FSD-based VR systems can provide a cost-efficient and flexible option for VR, especially when web-based remote testing is desired.

1.6.2 Studies comparing FSD- and HMD-VR

As mentioned in section 1.6.1, FSD- and HMD-based VR systems share many similarities but also have their distinctive advantages. Since VR can be used in many different fields that have different requirements (Cipresso et al., 2018; Slater & Sanchez-Vives, 2016), it is often worthwhile to compare the different technologies to decide which one would be more fitting for a given application. Next, studies that have compared the FSD- and HMD-based VR systems in learning outcomes, gaming, and traditional cognitive tasks with construct-driven design are reviewed.

Regarding learning outcomes, Makransky and others (2019) compared learning in HMD and FSD versions of a science lab simulation and found that the students reported having a stronger sense of presence in the HMD version while learning less and having a higher cognitive load based on electroencephalogram. Studying category learning, Barrett and others (2022) found no group differences in learning accuracy between HMD and two FSD conditions, but the participants in the HMD and 3D-stimuli+FSD conditions had slower reaction times than those in the 2D-stimuli+FSD condition. Ventura and colleagues (2019) examined the performance in memorizing items from a 360-degree picture in HMD and FSD conditions and found the performance to be better with the HMD. Next, Schloss and others (2021) compared learning about the functional anatomy of visual and auditory perceptual pathways either presented with HMD or FSD. Both technologies were effective platforms for teaching neuroanatomy with no differences between the devices, but participants reported enjoying the HMD format more than the FSD format. In a subsequent classroom implementation of the HMD version, students reported it as being effective in helping them visualize the perceptual systems.

As regards to commercial games, a recent meta-analysis found that the hardware (i.e., playing with an HMD and motion controller rather versus a FSD and non-motion controller) had a large effect on participants' sense of presence (Caroux, 2023). Several studies that compare FSD and HMD game versions have found that HMDs provide stronger sense of immersion and greater arousal of positive emotions (e.g., Pallavicini et al., 2019; Pallavicini & Pepe, 2019; Pallavicini et al., 2018; Tan et al., 2015), as well as higher user satisfaction (Shelstad et al., 2017). Regarding gaming performance, the results have been

mixed with some studies reporting no difference between the FSD and HMD versions (e.g., Pallavicini et al., 2019; Pallavicini & Pepe, 2019; Pallavicini et al., 2018; Shelstad et al., 2017) while other studies have reported higher performance in the FSD than HMD condition (Tan et al., 2015; Martel et al., 2015). As suggested by Pallavicini and colleagues (2019), the observed differences in the results could be due to changes in the hardware as the studies that have found superior performance in FSD condition (Tan et al., 2015; Martel et al., 2015) have included older HMD models. In addition, the game type and differences in the implementation of the controls could play some role here.

Several traditional cognitive tasks with a construct-driven design have first been implemented with the FSD and later adopted to the HMD. The original versions of these tasks have been non-immersive by nature, as they employ two-dimensional stimuli, unnaturalistic responses (e.g., using a keyboard or button box) and stimulus dynamics, and substantially diverge from looking realistic (Kourtesis & MacPherson, 2021). To better simulate everyday life, the HMD versions typically take use of the extended capabilities of the technology, for example, by embedding the task to a virtual scene (e.g., Armstrong et al., 2013; Brooks et al., 2017; Chang et al., 2020; Díaz-Orueta et al., 2014; Neguț et al., 2017) or by embedding distracting stimuli to the task (see, e.g., Neguț et al., 2017; Parsons et al., 2019). Studies have found correlations between the HMD and FSD versions suggesting convergent validity, for example, in the Stroop (Armstrong et al., 2013) and CPT tasks (e.g., Díaz-Orueta et al., 2014, see also the meta-analysis by Parsons et al., 2019). Typically, the sense of presence is evaluated to be stronger (e.g., Li et al., 2020) and the experience more positive with HMD than with FSD (e.g., Chang et al., 2020), but with older HMD hardware opposite results have also been reported (e.g., Brooks et al., 2017; Rand et al., 2009; for discussion on the differences between generations of HMD hardware on this matter, see Kourtesis et al., 2019a). Like with commercial games, some studies report equal task performance between the two platforms (Brooks et al., 2017), while others have found enhanced attentional performance with HMD (e.g., Li et al., 2020).

Taken as a whole, these studies of learning outcomes, commercial games, and cognitive tasks provide an important reference for the studies implemented either with the FSD or HMD. Depending on the task and hardware, the learning outcomes and other task performance may either be equal between the platforms (Barrett et al., 2022; Brooks et al., 2017; Schloss et al., 2021), superior with the FSD (Makransky et al., 2019; Martel et al., 2015; Tan et al., 2015) or superior with the HMD (Li et al., 2020; Ventura et al., 2019). The current HMD systems seem to yield higher sense of presence (Caroux, 2023; Makransky et al., 2019; Pallavicini et al., 2019; Pallavicini & Pepe, 2019; Pallavicini et al., 2018; Tan et al., 2015) and greater enjoyment/user satisfaction (Chang et al., 2020; Schloss et al., 2021; Shelstad et al., 2017) than the FSD, contrary to the results reported with some

earlier HMD models (Brooks et al., 2017; Rand et al., 2009; see also Kourtesis et al., 2019a). However, these studies do not address the implications of platform choice for function-led tasks that involve open-ended naturalistic scenarios from everyday life.

1.6.3 Function-led VR task in adults

As VR offers flexible and cost-efficient ways to create reproducible environments and allows different types of behavioral responses (e.g., movements of the eyes, head and body) to be measured accurately (see Campbell et al., 2009), it has been recognized as an ideal way of implementing function-led tasks with high ecological validity (Chan et al., 2008; Parsons, 2015; Parsons et al., 2017). Function-led naturalistic tasks mimic real-life situations and can be sensitive to cognitive impairments in situations where traditional tasks fail to do so (Cipresso et al., 2014; Shallice & Burgess, 1991). Potentially, they could also offer higher predictive value for everyday functions (Burgess et al., 2006; Chan et al., 2008; Parsons et al., 2017). As function-led paradigms can take full advantage of the new possibilities of immersive VR technologies, they have the potential to become the hallmark of VR-based cognition research (see Parsons et al., 2017). For adults, several studies with naturalistic VR tasks that simulate daily functions and activities already exist, both for FSDs and HMDs. The related findings are summarized below.

The contexts and situations simulated in FSD-based VR tasks include, for example, shopping (Canty et al., 2014; Cipresso et al., 2014; Grewe et al., 2014; Matheis et al., 2007; Parsons & Barnett, 2017; Rand et al., 2009; Raspelli et al., 2012; Ruse et al., 2014; Webb et al., 2021), food preparation (Allain et al., 2014; Besnard et al., 2016), running a library (Renison et al., 2012), and doing multiple chores in a city environment (Jovanovski et al., 2012). Also, virtual home (Sauzéon et al., 2012) and city environments (Plancher et al., 2012) have been mostly used to study memory performance in ecologically valid environments. In these studies, clinical groups, such as patients with Alzheimer's disease (Allain et al., 2014; Plancher et al., 2012), amnestic mild cognitive impairment (Plancher et al., 2012), traumatic brain injury (Besnard et al., 2016; Canty et al., 2014; Renison et al., 2012), stroke (Rand et al., 2009; Raspelli et al., 2012; Webb et al., 2021), schizophrenia (Ruse et al., 2014), Parkinson's disease (Cipresso et al., 2014) and focal epilepsy (Grewe et al., 2014) seem to perform worse than their controls on different measures, such as performance time, the number of errors/rule breaks, and learning performance.

Regarding HMD-based tasks, the simulated situations have included, for instance, shopping (Ouellet et al., 2018; Porffy et al., 2022) and food preparation (Barnett et al., 2021; Chicchi Giglioli et al., 2021). These studies have found that

the patients with a neurocognitive diagnosis (Barnett et al., 2021), alcohol use disorder (Chicchi Giglioli et al., 2021) or psychosis (Porffy et al., 2022) perform worse than their healthy controls on measures of memory and EF. Systems with HMDs have also been successfully employed to study memory function in ecologically valid task environments (Parsons & Rizzo, 2008). As a particularly comprehensive example of the potential of HMD systems, the Virtual Reality Everyday Assessment Lab (VR-EAL; Kourtesis et al., 2020; Kourtesis et al., 2021) follows a storyline with several scenes and functions that take place in several different environments (a bedroom, a kitchen, a garden, a car, a supermarket, a bakery, and a library).

To summarize, the literature of function-led VR tasks in adults is already sizable (see Neguț et al., 2016; Parsons, 2015; Parsons et al., 2017; Pieri et al., 2023), and new HMD-based research is being published at an accelerating pace (see, e.g., the review by Kim et al., 2021).

1.6.4 Function-led VR tasks in children

Contrasting the abundant function-led VR adult literature cited in section 1.6.3, the literature on respective child studies is scarce. Using non-immersive VR and a function-led task, it has been found that children with traumatic brain injury use more time and do more mistakes than their controls during a shopping task (Erez et al., 2013). Moreover, children with acquired brain injury perform worse than their controls on a task that requires multitasking and prospective planning during a birthday party preparation (Gilboa et al., 2019). Clancy and others (2010) used immersive VR to study road-crossing behavior and found children with ADHD to use lower safety margin, walk slower, underutilize the available gap in incoming traffic, show greater variability in road-crossing behavior, and be involved in twice as many collisions than their typically developing peers.

Instead of using function-led tasks, many HMD studies in children have focused on construct-driven paradigms, such as the CPT, that have been embedded to a virtual scene. Perhaps the most notable of the HMD-based CPTs has been the Virtual Classroom (Parsons et al., 2019; Parsons & Rizzo, 2019; Rizzo et al., 2009; for other HMD-based CPTs see the review by Pieri et al., 2023). It has been found that children with ADHD perform worse than their controls on the Virtual Classroom (Parsons et al., 2007). Adding ecologically valid distractors to the Virtual Classroom increases its ability to distinguish adolescents with/without ADHD (Adams et al., 2009), and methylphenidate decreases omission errors of children with ADHD in it (Mühlberger et al., 2020; Pollak et al., 2010). Furthermore, the Virtual Classroom seems more sensitive to subtle attention deficits related to sport-induced concussions than the traditional CPT (Nolin et al.,

2012), and children with neurofibromatosis type 1 do more commission and omission errors than their controls (Gilboa et al., 2011).

Importantly, there have been no child studies that would employ an immersive (i.e., HMD-based) virtual environment and use function-led task design to probe the children's abilities to perform varied everyday chores in a life-like environment.

1.7 General prerequisites of psychometric measures

All psychometric measures and tests, whether designed in a construct-driven or function-led manner, should meet certain criteria to be accepted for their intended purpose. Among these prerequisites are *validity*, which has already been discussed in the form of ecological validity, and *reliability*. Broadly speaking, validity refers to the extent to which an instrument measures the intended construct, and reliability refers to how accurately it does so (Raykov & Marcoulides, 2011). Validity and reliability of a measure are often evaluated quantitatively. This can be done, for example, by using different correlation coefficients, such as the Pearson correlation coefficient (Raykov & Marcoulides, 2011), different intra-class correlation coefficients (see, e.g., Koo & Li, 2015), or concepts related to signal-detection theory, such as *receiver operating characteristic* (ROC) curve and *area under the curve* (AUC; see, e.g., Stanislaw & Todorov, 1999). Both validity and reliability have different forms. Next, the forms used in the present Thesis are defined.

1.7.1 Concurrent, predictive, and discriminant validity

Concurrent and *predictive validity* can be regarded as two types of criterion validity (Cronbach & Meehl, 1955). In both, the researcher is interested in some criterion which (s)he wishes to predict (Cronbach & Meehl, 1955).

To estimate concurrent validity, the researcher should obtain the test and the criterion at the same time (Cronbach & Meehl, 1955). For example, if a test that seeks to measure ADHD symptoms is conducted while asking the parent to assess the same symptoms by filling out a rating questionnaire, and a correlation is calculated between the test result and the rating questionnaire, this correlation can be regarded as reflecting concurrent validity of the test while using the rating questionnaire as the criterion (Nichols & Waschbusch, 2003). This definition of concurrent validity can be especially suited for function-led measures, as in them the target phenomenon is defined as on the functional level (Burgess et al., 2006), and ADHD rating scales are enquiries about observations collected during daily activities, that is, functions, of the individual. In contrast, construct-driven

measures are usually seen as targeting hypothesized cognitive constructs, which are on a different level of explanation (Burgess et al., 2006).

Predictive validity differs from concurrent validity in that the criterion and the test result are not obtained simultaneously. Usually, the criterion is collected after the test result (see Cronbach & Meehl, 1955), but can also be obtained before the test result. For example, if a test shows group differences between a clinical group and control group, it can be regarded as having predictive validity to that clinical condition.

In this Thesis, following the paper by Nichols and Waschbusch (2003), discriminant validity refers to a measure's ability to successfully classify participants on some criterion based on their measured performance. For example, this criterion could be diagnostic status such as whether a participant has ADHD or not. To summarize, concurrent, predictive and discriminant validity can all be used to evaluate if a measure successfully quantifies differences related to a given criterion, such as ADHD.

1.7.2 Reliability as internal consistency and test-retest stability

In classical test theory, each measurement or observed score is regarded as consisting of two parts, the measurement error and true score that is of interest to evaluate (Raykov, Marcoulides, 2011). As only the observed score can be directly measured, it is important to evaluate, how closely it corresponds to the true score that cannot be measured. Reliability is an index that indicates how much information about the true score is contained in the observed score (Raykov & Marcoulides, 2011). This corresponds to the correlation between the two and can only be estimated, not accurately measured (McDonald, 1999; Raykov & Marcoulides, 2011). Establishing reliability is essential not only for judging the consistency of a measure, but also for considering its validity, as no validity coefficient can be interpreted without some appropriate estimate of the magnitude of measurement error (i.e., reliability; Cronbach, 1951). Reliability can be estimated using several ways, of which *internal consistency measures* and *test-retest stability* are covered here.

Internal consistency indicates how closely the individual items of a measure are related and can be acquired by calculating the correlation between different subsets of items within the measure (Cronbach, 1951). Of the internal consistency measures and reliability measures in general, the Cronbach's alpha has been the most widely used in psychological research (Hogan et al., 2000; see also Trizano-Hermosilla and Alvarado, 2016). Cronbach's alpha is the mean of all split-half reliabilities and assumes tau-equivalence, that is, that all test items have equal factor loadings on the latent variable being measured (Cronbach, 1951). Cronbach's alpha has been criticized and its use questioned especially in situations

where the test items differ in quality or the distributions are skewed (Trizano-Hermosilla and Alvarado, 2016; McNeish, 2018). When the tau-equivalence is violated Cronbach's alpha will produce be an underestimation of true reliability up to 11.1 % (Green and Yang, 2009). Because of the shortcomings of Cronbach's alpha, many other internal consistency measures have been developed, such as McDonald's omega (McDonald, 1999; see also McNeish, 2018; Trizano-Hermosilla and Alvarado, 2016). Nevertheless, Cronbach's alpha remains the most popular option for determining the internal consistency of a measure (McNeish, 2018).

Test-retest reliability (stability) can be established by administering the same test or measure on two different occasions and examining the correlation between observed scores to evaluate stability over time (Raykov & Marcoulides, 2011). Some researchers prefer referring to the correlation between two occasions of the same measure as test-retest stability instead of reliability, as the temporary changes in participants' true scores affect these correlations (Raykov & Marcoulides, 2011; Cronbach, 1951). For example, Cronbach (1951) states that a retest after an interval with an identical test indicates how stable scores are and should therefore be called a coefficient of stability. Regardless of whether the correlation of two instances of the same measure is referred as test-retest reliability or stability, it is important to be established, especially if a measure is intended for repeated assessments, which is common in clinical practice (Lo et al., 2012). After Cronbach's alpha, test-retest stability is the second most common reliability coefficient in psychological research (Hogan et al., 2000).

2 Aims of the study

The overarching aim of the present Thesis was to develop and apply a new VR task for the assessment of children's goal-directed behavior and ADHD symptoms in real-life contexts. The development was inspired by naturalistic PM studies (e.g., Rendell & Craik, 2000), studies carried out in real-life environments, such as the MET (Shallice & Burgess, 1991), and computerized tasks that simulate real-world environments (e.g., Rand, 2009; Cipresso et al, 2014). This novel VR task, named *Executive Performance in Everyday LIving* (EPELI)[1], was designed with equal contributions by Matti Laine, Juha Salmitaival (aka Salmi), and Erik Seesjärvi.

Following the rationale of function-led task design described in the introduction, EPELI was intended as a more versatile and ecologically valid instrument for the assessment of goal-directed behavior, which requires a diverse set of cognitive abilities, including EF and PM. These functions have been found to be impaired in ADHD (e.g., Barkley, 2012; Kliegel et al., 2006). Furthermore, it was presumed that simulating everyday situations would allow the measurement of other behavioral aspects as well, such as hyperactivity and impulsivity, which are key characteristics in ADHD (Faraone et al., 2021). The freedom to interact with an engaging open-ended realistic environment creates an immersive illusion of real life (Bohil et al., 2011; Slater, 2018), which was expected to prompt typical ADHD-related behaviors, such as impulsive actions directed towards attractive but task-irrelevant stimuli. The function-led approach was chosen to overcome some of the key limitations of simplified, "construct-driven" laboratory tasks (e.g., Barkley et al., 2012; Burgess et al., 2006; Chaytor et al., 2006; Dawson & Marcotte, 2017; Miller & Barr, 2017).

To our knowledge, EPELI is the first VR-based task for children that requires them to carry out various everyday chores while planning their movement around the virtual environment, monitoring the time, and avoiding getting distracted by irrelevant objects, sounds, or events. EPELI was first implemented with HMD technology (i.e., HMD-EPELI), which was used in Studies I, II, and III. In Study

[1] See https://aalto.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=3eb4836f-1238-4f27-853a-ad3700745b31 for an example video of EPELI.

IV, another EPELI version, which uses the widely available FSD technology, was used (i.e., FSD-EPELI). Next, the specific aims of each Study are reviewed.

2.1 Specific aims of Studies I–IV

In Study I, the main aim was to test the predictive, discriminant and concurrent validity of EPELI using groups of children with/without ADHD. To do so, we operationalized a diverse set of measures targeted at quantifying goal-directed behavior and ADHD symptoms. The study was preregistered (https://aspredicted.org/qk7js.pdf). As regards predictive, discriminant, and concurrent validity, it was hypothesized that: i) the ADHD children's task efficacy would be lower than that of the controls, reflecting attentional-executive function problems; ii) that the movement trajectories when moving around the environment would be longer and the motion sensors would detect higher levels of motion with the ADHD children than the controls, indicating hyperactivity; and iii) because of impulsivity, the ADHD children would perform more actions than the controls. Based on previous VR research, these differences were expected to be pronounced in scenarios with more distractors (Neguț et al., 2017; Parsons et al., 2007). We also expected that the ADHD participants would show higher variability in the EPELI measures (Sonuga-Barke & Castellanos, 2007) and that their performance would not improve during EPELI, unlike that of the typically developing controls. Regarding concurrent validity, it was expected that the EPELI measures would correlate with parent evaluations of ADHD symptoms and EF deficits, as EPELI scenarios simulate everyday situations akin to those in which these symptoms and deficits manifest themselves.

In Study II, the key aims were to replicate the findings of Study I concerning the predictive and discriminant validity of EPELI using another data set comprising groups of children with/without ADHD and to collect additional eye movement data for further analyses. It has hypothesized in the preregistration (https://tinyurl.com/yck2y7u2) that, besides replication of the findings of Study I, (i) eye movement data would further improve EPELI's discriminant validity (i.e., prediction of group status); (ii) salient objects would attract the gaze of ADHD children more effectively than that of typically developing control children; (iii) children with ADHD would focus less efficiently on relevant stimuli than the controls; and (iv) the ADHD group would exhibit a greater number of switches between ambient and focal processing due to fluctuating attention.

The main aim of Study III was to examine several key properties of EPELI in typically developing children, and thus evaluate its applicability to further research and clinical work. The key properties examined included psychometric features of internal consistency reflecting reliability and the two types of ecological validity, namely veridicality and verisimilitude. Furthermore, the possible effects

of age, gender, gaming background, previous familiarity with the task contents, HMD type, and verbal recall ability on EPELI measures were tested. Finally, the associations between parent-rated EF problems and EPELI measures reported in Study I were reinvestigated in greater detail: that is, we probed whether EPELI performance would correlate either only with problems in behavioral regulation, metacognitive skills, or both. In addition to the five EPELI measures employed in Studies I and II, another three EPELI measures specific to PM were used.

In Study IV, the primary aim was to compare immersive and non-immersive implementations of EPELI. The first version, HMD-EPELI, was used in Studies I–III and implemented with HMD technology. The second version, FSD-EPELI, was developed for Study IV and is based on the FSD technology, which is widely available as the standard display technology in laptop and desktop computers. All participants performed both versions in an order that was counter-balanced between the participants. To see if parent-supervised remote testing would yield results comparable to those obtained in laboratory settings, we asked half of the participants to perform FSD-EPELI at home, with their own computers. The more specific aims of Study IV were: (i) to test possible differences in the eight EPELI measures between the two versions and learning effects between the first and second assessment; (ii) to test possible differences in subjective experience ratings between the two versions; (iii) to examine possible differences between the EPELI measures and subjective ratings between experimenter-supervised laboratory testing and parent-supervised home testing; (iv) to reinspect the associations between parent-rated difficulties of executive function and EPELI, which had been reported in Studies I and III using HMD-EPELI and also with FSD-EPELI; and (v) to evaluate the inter-version correlations and test-retest stability of EPELI.

3 Methods

3.1 Participants

A total of 85 children with ADHD and 146 typically developing children took part in the studies of this Dissertation. After exclusion and propensity matching procedures (see below), the final samples included data from 74 children with ADHD and 123 typically developing children, some of them taking part in more than one study (see Figure 1).

All participants were school-aged children from southern Finland. The typically developing children (Studies I–IV) were recruited from schools in Espoo and Kirkkonummi by inviting them to participate after a lecture where they had been informed about the given study and by sending recruitment letters to their parents via schools' electronic message board Wilma. The eligibility of each typically developing child was initially checked on the first contact (phone call or email) with the parent and later controlled from the parent questionnaires, where the parents listed possible medical diagnoses of their child. The participants with ADHD (Studies I–II) were recruited via child psychiatric units of Helsinki University Hospital, Finnish ADHD Foundation contact channels, the Espoo City Child Psychiatric Unit, the Vantaa Family Counselling Unit, and a private clinic in Espoo (ProNeuron LTD). Also for the ADHD participants, the eligibility was initially probed on the first contact with the parent. For children with ADHD, all diagnoses were confirmed from medical records and the other criteria from the parent questionnaires. Before participating, all children and their parents gave their informed consent according to the declaration of Helsinki. All studies were approved by the Ethics Committee of the Helsinki University Hospital. The children received either two (Studies I–III) or four (Study IV) movie tickets as a compensation for participating.

Study I included two groups, one with children diagnosed with ADHD and another with typically developing controls. For the ADHD group, the inclusion criteria were (i) ADHD diagnosis with predominantly hyperactive/impulsive or combined inattention and hyperactive-impulsive subtype (F90) set by a licensed physician using the ICD-10 criteria (World Health Organization, 2016), (ii) age of 9–12 years when recruited, and (iii) native language Finnish. The exclusion criteria were (i) any diseases of the nervous system (ICD-10, G00–G99) and (ii)

any additional mental and behavioral disorders (F00–F99) except F93 (Emotional disorder with onset specific to childhood) and F98 (Unspecified behavioral and emotional disorder), which were allowed as secondary diagnoses as being common comorbidities in ADHD. For the control group, the criteria were the same, except that the exclusion criteria included any mental or behavioral disorders (F00–F99) and the need for special support at school. In total, 47 children with ADHD and 68 typically developing controls took part in the study. Nine children with ADHD and 17 control participants had to be excluded due to not fitting the inclusion criteria or because of incomplete EPELI data. To select a matching control group for the ADHD group from the remaining 51 control participants, propensity matching with age, gender, parental education, and familial income as the matching variables was performed using R package MatchIt (Ho et al., 2011). The final sample consisted of 38 ADHD and 38 control participants with no group differences in age, handedness, gender, parental education, familial income, or reasoning abilities (see Table 2). In the ADHD group, 30 out of 38 children had a medication to alleviate their ADHD symptoms, but they took part in the study unmedicated (24-hour washout period).

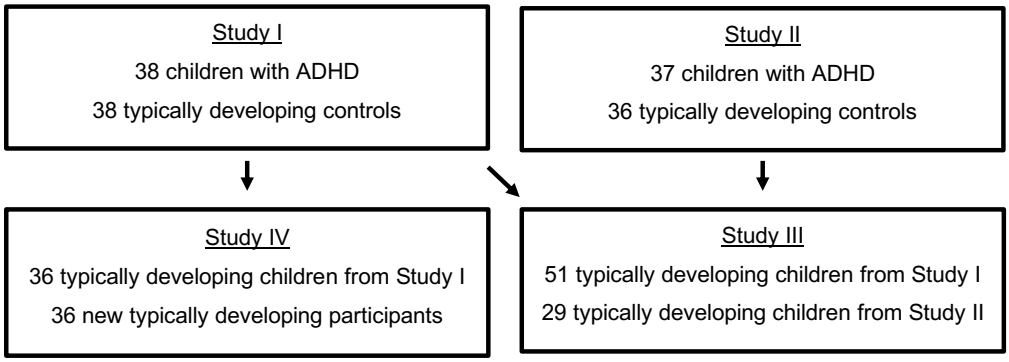
In Study II, the sample consisted of groups of ADHD children and typically developing controls. For both groups, the inclusion and exclusion criteria were the same as in Study I, except that children with predominantly inattentive subtype of ADHD were also allowed to take part in the study. Before data analysis, four children from the control group had to be excluded due to technical problems. Also, one control participant was excluded because of difficulties with Finnish language due to significant time living abroad and one participant with ADHD because of dropping out during the measurements. The final sample comprised 37 children with ADHD and 36 typically developing controls. The groups were matching on age and handedness, but control group had higher familial income and performed better on reasoning tasks (see Table 2). In the ADHD group, 34 children had a prescription for their ADHD symptoms, but they took part in the study unmedicated (24-hour washout period). In MINI-KID interview (Sheehan et al., 1998), 12 children with ADHD met the criteria for some other psychiatric condition.

In Study III, the data of the typically developing children, who had participated in Studies I and II, was used. After the exclusion of 20 participants due to incomplete EPELI data and outlier analyses, the final sample comprised 77 children (see Table 3).

Study IV included 101 typically developing children, and the inclusion criteria were the same as in the Study III. The control participants from the Study I were invited to this study, and new participants were recruited by sending new advertisement letters to parents via schools' electronic message board Wilma and giving brief educational lectures at schools and online, during which the

opportunity to take part in the study was mentioned. The data for one of the sessions (see 3.3.4) was missing from 29 children because of dropping out of the study after the first sessions or due to technical issues. The final sample included 72 typically developing children (see Table 3).

Figure 1 A chart that shows the number of participants in the **final** sample of each study and illustrates the relationships between the samples.



3.2 The EPELI task

The following sections describe the development and contents of HMD-EPELI and FSD-EPELI, which were both implemented by Peili Vision company (https://peilivision.fi). After that, the measures operationalized in the studies of the present Thesis are described.

3.2.1 The development and task contents of EPELI

After the initial conceptualization, the development was started by choosing a suitable virtual environment for the task. A typical home environment was chosen among several alternatives (e.g., a classroom, a grocery store, and a cafeteria) as the concept of home was evaluated to be familiar enough for all children to ensure an unbiased evaluation, and a home environment allows many kinds of typical real-life scenarios to be implemented. The apartment was designed to include multiple rooms: a children's room, living room, kitchen, open adult bedroom, utility room, bathroom, and a balcony that could be seen through windows but not be accessed. The multiroom layout allowed the scenarios to require moving and thus the planning of movement around a bigger, non-unified area. The floor plan of the apartment, which was *not* shown to the participants at any point, can be seen in Figure 2.

Figure 2 The floor plan of the apartment in EPELI (**not** shown to the participants). Copyright Peili Vision, 2022. Adapted with permission from Study III.

After choosing the environment, instructions for several everyday scenarios were written. The scenarios were chosen as situations where difficulties of goal-directed behavior may occur in real life and as having themes that would be easily recognizable for school-aged children (e.g., going to school, coming back from a hobby, and preparing a meal). During piloting phases, some difficult or unambiguous words causing confusion were changed to make the instructions easy to understand. Furthermore, the participants of Study I were asked to name or recognize by name a group of key items and objects used in EPELI, which they all were able to do (see Supplementary Table in Study I). A cartoon dragon character named Laura Lohikäärme (Laura the Dragon) was chosen to give the instructions to the children. An animal cartoon character was chosen as being more entertaining for the children than a human character and to avoid the uncanny valley effect, which refers to a situation where an artificial agent is very lifelike but still clearly recognizable as non-human, which can feel disturbing (Mori et al., 2012).

After several piloting iterations, a final set of scenarios were chosen. This final set includes a practice session and 13 scenarios. In the practice session, Laura Lohikäärme helps the participant to become familiar with the environment and navigate in it, interact with the objects, and monitor the time. The practice session has no time limit and takes around a couple of minutes. The 13 scenarios comprise

a total of 70 subtasks, of which 52 can be completed at any time, while 13 are to be performed at a given time (TBPM subtasks) and 5 after an external cue (a given sound, such as a signal from the dish-washer, doorbell or cell phone tone; EBPM subtasks).

Each EPELI scenario starts with *an encoding phase*, during which Laura Lohikäärme orally gives the instructions of the scenario to the child. The instructions start by mentioning the theme (e.g., ”you are just about to leave for the school”) and continue by listing the subtasks (e.g., ”turn off the tap that your father forgot to close”, ”call your mother at two o’clock”, ”take your backpack to your room”). Depending on the scenario, the encoding phase includes four to six subtasks covered by 30–66 words and takes 22–47 seconds.

The encoding phase is followed by *an execution phase*, which lasts a maximum of 90 seconds but ends earlier if all subtasks are performed before that. During the execution phase, the child should perform the given tasks by navigating the environment and interacting with relevant objects. The subtasks need be executed from memory, as the instructions are not repeated. During the development, some experimentation was also made with versions in which the child could ask the remaining instructions to be repeated and the maximum duration of the execution phase was longer. However, the children used the possibility to check the instructions very differently. Therefore, to ensure the uniformity of the task to all children, this option was later removed, the maximum duration was set at 90 seconds, and the number of scenarios was increased. The maximum duration of 90 seconds allows the given subtasks to be performed even at an easy pace, if the participant can successfully engage in effective goal-directed behavior without being distracted with non-irrelevant activities and stimuli and remember the remaining subtasks. The participant is asked to perform the subtasks in the given order, except for the EBPM and TBPM subtasks. However, the completion order does not influence the scoring (see 3.2.5 EPELI measures).

The maximum duration of the final EPELI version with a practice session and 13 scenarios was found to be around 35 minutes, but on average, the children performed it somewhat faster (on average in 27 minutes in Study IV). As the research group had no previous experience about how long VR sessions are suitable for school-aged children, the scenarios were made independent so that it would have been possible to shorthen the overall duration of EPELI by removing some scenarios. However, if was found that a vast majority of children could perform this the set of 13 scenarios nonstop with the absence of any sickness symptoms. The very few children who reported any nausea after the task were nevertheless able to complete the whole task, and some of them spontaneously commented, that the symptoms they reported were not related to VR but had been present already before EPELI. In all studies of this Thesis, the order of the

scenarios was counterbalanced between the participants, that is, every other participant performed them in reversed order.

To emulate stimulus-rich everyday environments, the possibility to embed the scenarios with additional distractors was also introduced. These included (i) auditory distractors, such as traffic noises, music coming from the radio, dog barking, and child coughing, and (ii) audiovisual distractors, like a fly buzzing nearby the character, a TV with several channels, and a tap left running. In all studies, the distractors were used in seven (for participants performing the scenarios in forward order) or six (for participants conducting the sevarios in reverse order) scenarios. These scenarios also contained more task-irrelevant objects. The distractions were counterbalanced between the participants the same way as the order of the scenarios, which meant that all scenarios were played both with and without distractors, but by different participants. In the distracted scenarios, the distractors were on during the whole scenario except for the TV, radio and running tap, which the participant could turn off.

Time monitoring was given some special consideration. As some children may not be familiar with analog watch with moving hands showing the hours, minutes, and seconds, it was decided that a simplified version would be used. The simplified version has numbers from zero to three and only one hand that does a full circle in 60 seconds (see Figure 3). While the hand is moving, the clock face turns gradually red, like a reversed time timer. During piloting, it was confirmed that all test children could easily learn how the clock works the way it was explained by the dragon character. To follow the time monitoring behavior of the children, it was decided that the watch should be hidden unless the children purposefully take an action to view it, as is often done in PM paradigms (e.g., Kerns, 2000; Zuber et al., 2021).

For walking around the apartment, several options were considered. Even though the current HMDs like Pico Neo 3 Pro track the movement of the headset in physical space and thus allow navigation to be performed by simply walking in physical space, this implementation was evaluated to be an unpractical especially regarding portability. The navigation was therefore decided to be implemented via clickable waypoint circles on the floor. Clicking a waypoint results the participant to teleport to the corresponding location in the virtual space. First, this movement was animated, that is, the participant could see him/herself gliding to the new location. After some pilot participants complained that the gliding caused them slight nausea, it was substitued with instant teleporting, which eliminated the problem. For the reasons mentioned in section 3.2.2, all children performed EPELI in a sitting position, even though using a standing position is also possible with HMDs.

3.2.2 HMD-EPELI and related hardware

Before starting the actual data collection in Study I, we conducted several pilot measurements to decide, whether the task should be performed in a standing or seated position. Especially younger children with no prior VR experience had problems when playing in a standing position. For example, they tried to reach a table or wall to lean to and reported mild adverse effects, such as feeling dizzy. We therefore instructed the children to play in a seated position, even though a standing position could provide better sensorimotor contingency. This was to assure safety and avoid problems that could potentially compromise data quality, such as cybersickness.

To look around in HMD-EPELI, the participants can rotate their head and spin with the chair. To interact with objects, they can point objects with a ray coming from the virtual hand controller object and push a button. With Oculus Go, the button is pressed with the thumb and with Pico Neo using the index finger. Movable objects can be picked up by pointing at them and pressing the button and released by pointing at a desired location and pushing the button again. To check the time, the participant needs to rise the controller slightly and look at it, like checking the time from a wristwatch. The drums in the virtual environment can be played by swinging the hand controller at them. This is required in only one scenario but can spontaneously be done in others. Navigating is performed by pointing at the desired waypoint circle and pressing the button. This one-button approach was chosen to make the interface easy to use for participants who may have only limited or no previous experience with HMD technology.

In Studies I and IV, the HMD hardware was Oculus Go (2560 x 1440 resolution, 16:9 aspect ratio, 60/72 Hz refresh rate, and 101-degree horizontal field of view). In Study II, Pico Neo 2 Eye goggles (3840 x 2160 resolution, 16:9 aspect ratio, 75 Hz refresh rate, and 101-degree horizontal field of view) were used. Eye tracking was done using the inbuilt Tobii eye tracker (90 Hz recording rate, 0.5 degree stated system accuracy). Study III used data pooled from Studies I and II, which allowed the two HMD models to be compared.

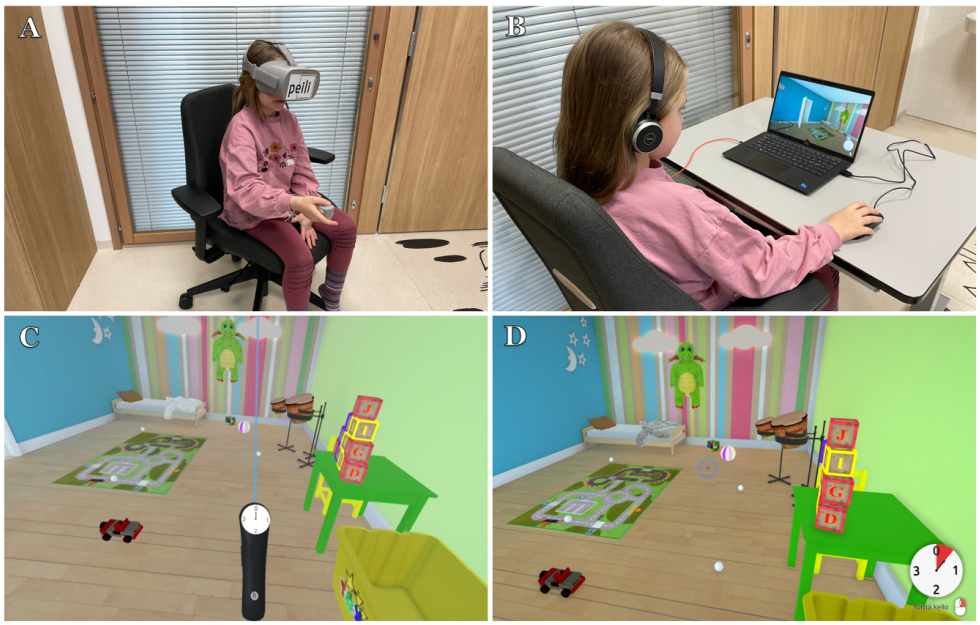


Figure 3 Pictures and screenshots from EPELI. (A), a child during HMD-EPELI. (B), the same child performing FSD-EPELI. (C), an HMD-EPELI screenshot that shows the virtual hand controller with the clock. (D), an FSD-EPELI screenshot that shows the clock in the lower right corner of the screen and the crosshairs in the middle of the screen. Adapted with permission from Study IV.

The two HMD models differ in the way they track participant motion: the Oculus Go tracks only rotation of the headset and controller, and the movement of head and hand in virtual space is estimated from the rotation (three degrees of freedom tracking), while the Pico Neo 2 Eye tracks both position and rotation, therefore moving the head or hand results in similar movement in the virtual space (six degrees of freedom tracking). For audio, both models have integrated speakers located near (~ 3 cm) each ear. In all studies, the volume was kept fixed at a predetermined level that allowed the children to hear the instructions easily. The examiner was encouraged to adjust the loudness if the participant could not hear the speech of the dragon during the practice session, but this was not necessary for any of the participants. In the Presence questionnaire that was presented after EPELI (see section 3.3.1), all participants reported that they had heard the instructions easily. The experimenter launched EPELI and could follow the gameplay using a Samsung Galaxy Tab S2/3 tablet.

3.2.3 FSD-EPELI and related hardware

FSD-EPELI was developed for Study IV. The key differences to the HMD version are as follows: (i) in FSD-EPELI, the participant uses a mouse/trackpad to change

the direction of the view instead of rotating the head, (ii) the FOV is markedly smaller (depending on the screen size and viewing distance, approximately 25–60 degrees as measured horizontally), (iii) the view of the surrounding physical environment is not blocked as with HMDs, (iv) the interaction with the objects is done by rotating the view until the desired object is located in the crosshairs in the middle of the screen, which means that unlike in the HMD version, the participant cannot interact with the objects independently from the direction of the view, (v) the watch is viewed by clicking the right mouse/trackpad button, which makes the white circle located at the lower right corner of the screen to reveal a clock (see Figure 3).

The hardware used for FSD-EPELI in Study IV was as follows. The participants who performed FSD-EPELI supervised by an experimenter used a Dell Latitude 7400 laptop computer (Inter i5-8365 CPU @ 1,6 MHz, 16 GB of RAM, Intel UHD Graphics 620 GPU, 14-inch screen, 1920 x 1080 resolution, 60 Hz refresh rate, 16:9 aspect ratio, Windows 10 OS), a Logitech M100 mouse, and Jabra Evolve 40 MS headphones. The participants who played FSD-EPELI at home used various typical laptop/desktop configurations. The screen size was 12–32 inches with the most popular options being 14 inches (four participants) and 24 inches (five participants). As the viewing distance was not measured, the exact FOV is not known. It can be approximated that when a typical viewing distance of 60 cm was used, the horizontal FOV varied between 25–60 degrees, meaning it was markedly narrower than that of the HMD (101 degrees). Of the participants who performed EPELI at home, 25 used a mouse, four a trackpad, and six failed to report this. For audio, 18 participants used headphones, 11 conventional speakers, and six did not report this. In the FSD version, the correct loudness level was determined individually for each participant by first finding the threshold level of hearing and then setting the volume to a comfortable level above it. If needed, the children could manually adjust the volume before starting EPELI.

3.2.4 The EPELI measures

During the development of EPELI, several indices and measures were first operationalized. In data analyses of Study I, 14 measures were chosen as reflecting different aspects of goal-directed behavior and the key features of ADHD symptomatology, namely attentional-executive function problems, impulsivity, and hyperactivity (see Study I, Supplementary Table 1). The final number of measures was reduced to eight by examining their pairwise correlations in the sample of typically developing children and dropping out one measure from each pair with a correlation of .85 or more. This final set of measures included Total score, Task efficacy, Navigation efficacy, Controller motion, Total actions, TBPM score, Clock checks, and EBPM score.

Total score (the number of correctly performed subtasks) taps the general prowess in executing multistage goals in a naturalistic condition. Being the number of correctly remembered and executed target items, it is considered to have a strong memory component. This interpretation is supported by the correlation between Total score and the Instruction recall task (r = .49 in Study I), where participants verbally repeat instructions similar than those given in EPELI. However, unlike the Instruction recall task, achieving high Total score also requires ability to plan and execute the given tasks.

Task efficacy measures how well the participant can selectively focus on executing the relevant goals. It is calculated as the percentage of relevant actions, that is, actions that where needed for the completion of the correctly performed subtasks, out of all actions excluding clicks on the waypoints that enable navigating around in the environment. As EPELI environment includes a wealth of irrelevant stimuli that can be distractive or attractive, efficient performance requires ability to inhibit irrelevant stimuli that may capture one's attention (e.g., a toy object that would be nice to play with) or cause interference (e.g., noticing and grabbing another fruit when the task would be to eat an orange). Thus, the definition of task efficacy comes close to a typical definition of selective attention, that is the ability to selectively process the relevant stimuli and ignore the irrelevant stimuli. However, unlike in typical experimental cognitive tasks, the irrelevant stimuli in EPELI are designed so that they would be rewarding and appealing to act upon.

Navigation efficacy reflects the economy of walking movement in the task environment. It is calculated by dividing the Total score by distance covered, which includes the distance walked and the distance to each manipulated object. Efficient navigation requires planning, strategy, and inhibition (i.e., avoiding any additional movements or operations and focusing on the relevant tasks). As Total score, Navigation efficacy is correlated with the Instruction recall task (r = .51 correlation in Study I).

Controller motion is a general measure of motor activity. It is the amount of angular controller movement in degrees and thus indicates the hand movement during the task. Only the movement during task execution phase of each scenario is included. While neither a low nor high level of controller motion is not necessarily an indication of good or bad performance, extreme values may be indicative of neuropsychiatric symptoms, such as hyperactivity symptoms in ADHD.

Total actions measure reflects the total activity with the VR environment. It includes the number of clicks during the task execution (also those used for navigating the environment and those that are not targeted at any object), the number of times hitting the drums in the child's room by swinging the controller, and the clicks done during the instruction phase of each task scenario. Like Task

efficacy, it reflects the amount of irrelevant activity, but without considering the overall task performance. Therefore, it captures the tendency to actively interact with different types of stimuli in the environment. As many objects in EPELI are designed to be attractive and tempting for children, Total actions is expected to reflect impulsive behaviors.

TBPM score reflects time-based prospective memory ability, that is, the accuracy in performing tasks in designated time. It is the number of time-based subtasks executed within ± 10 seconds from the target time. TBPM measures are widely used in PM research and supposed to reflect 'when' aspect in remembering to perform everyday life tasks (e.g., when to take food out of the oven and when to leave for school).

Clock checks reflects active time-monitoring behavior and is the number of times when the clock has been viewed. Like in real life, checking the time too often can be disadvantageous to the tasks at hand, but doing so at well-planned times to optimize the use of time (e.g., do I have time to go take my backpack to my room before I need to turn off the oven) can be important for good task performance, especially regarding time-dependent tasks.

EBPM score measures PM ability in the form of responsiveness to external memory cues. It comprises the number of event-based subtasks to be performed within 10 seconds from the start of the cue. External cue processing is one of the core processes in PM, meaning that seeing objects or hearing sounds related to a PM task often triggers us to recall what we were supposed to do. EBPM differs from TBPM in that no strategic monitoring of time is needed.

3.3 Procedure and collected data

The tasks and questionnaires in each Study are summarized below. For more details, see the corresponding section in the original publication of each Study.

3.3.1 Study I

In Study I, all participants performed the HMD version of EPELI. After EPELI, they orally answered a shortened version of the Presence Questionnaire 3.0 (Witmer et al., 2005), the Simulator Sickness Questionnaire (Kennedy et al., 1993), a gaming experience questionnaire, and an object familiarity questionnaire (see Study I Supplementary Tables 5 and 6). In the Presence Questionnaire, the children answered questions regarding the sense of presence during EPELI. Their possible cybersickness symptoms were probed with the Simulator Sickness Questionnaire and familiarity with computer, console, cell phone and VR games with the gaming experience questionnaire. The object familiarity questionnaire was used to ensure that the children were familiar with and could name the

objects used in EPELI. Also, to probe the children's familiarity with the task contents in EPELI they were asked "From a scale of 1 to 7, how much have you performed similar tasks in real life?".

The conventional neuropsychological tasks included the Similarities, Matrix reasoning and Digit span subtests from the Finnish version of the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003), the CPT (Rosvold et al., 1956), *Simple Reaction Task* (SRT; see Psycho-Motor Vigilance Task in Wilson et al., 2010), Cruiser task (Kliegel et al., 2013; see also CyberCruiser in Kerns & Price, 2001), the Clock task (see the Finger-snapping task in Kerns & Price, 2021), *Heidelberger Exekutivfunktionsdiagnostikum task* (HEXE; Kliegel et al., 2006), and *the Frogs and Cherries task* (F&C; see Dots & Triangles in Zuber et al., 2019). They also performed the Instruction recall task (aka the Repetition task), in which they had to orally repeat instruction akin to those heard during EPELI. For a more detailed description, see Supplementary Methods of Study I.

The parent questionnaires included the ADHD-RS (DuPaul, 1998), BRIEF (Gioia et al., 2000), *Child Behavior Checklist* (CBCL; Achenbach, 1991), and *Executive Questionnaire of Everyday LIfe* (EQELI; see Study I Supplementary Table 4). From ADHD-RS, the total score (i.e., sum of all questions, range 0–56) was selected as the dependent variable. In the BRIEF, the total score (i.e., *General Executive Composite* aka GEC; range 72–216) was used as the dependent variable. To gather information about problems in specific life scenarios (e.g., cleaning the house, preparing for school) presented in EPELI, we designed a new questionnaire, EQELI, from which the total score (i.e., sum of all questions, range 0–100) was used as the dependent variable. The CBCL was used to screen overall psychiatric symptoms, and total internalizing (range 0–62) and externalizing (range 0–64) symptoms were reported as background information. The participants with ADHD and their caregivers also took part in the MINI-KID interview (see 3.1 Participants).

3.3.2 Study II

In Study II, the participants performed HMD-EPELI and answered the same oral questionnaires than in Study I (see 3.3.1). They also performed the Instruction recall task, a VR visual search task (Shoot the Target), and the Similarities and Matrix Reasoning subtests from WISC-IV (Wechsler, 2003). In Shoot the Target task, the children were asked to search for target objects appearing in the visual space in front of them and shoot those targets by orienting their gaze to them (for details, see Supplementary Materials in Study II). The analyzed measures included Total Score (correctly shot targets out of the number of presented targets plus false alarms), Head Rotation Angle (sum of head rotation in degrees during the task), and Head Rotation Speed (average head angular speed). The parent

questionnaires and their dependent variables were the same as in Study I. Like in Study I, the participants with ADHD took part in the MINI-KID interview with their parents.

3.3.3 Study III

The data of Study III was drawn from the samples of Studies I and II. The dependent variables of each task and questionnaire were as described above, with the exception that from the BRIEF, three different dependent variables were used. The BRIEF questionnaire includes eight clinical scales that form two broad indexes, namely the *Behavioral Regulation Index* (BRI) and *Metacognition Index* (MI). When summed together, the BRI and MI form the overall score GEC (see section 3.2.1). In Study III, the raw BRI and MI scores were used as dependent variables in addition to the GEC used in Studies I and II.

3.3.4 Study IV

Study IV included two measurement sessions, one with HMD-EPELI and the other with FSD-EPELI. All HMD-EPELI sessions and 37 of the FSD-EPELI sessions were conducted either in laboratory at Aalto University or a comparable school room, and the remaining 35 FSD-EPELI sessions at home. In laboratory and school, the children were assisted and supervised by one of the researchers, and at home by a parent. The order of the sessions was counter-balanced between the participants, and the time between the sessions was 3.3 to 10.5 months. There was no difference in the average time between the sessions between the groups who performed the HMD session or FSD session first, as the average time between the sessions was 220 days for the HMD-first group and 208 days for the FSD-first group ($t(69.34) = 0.980$, $p = .33$). However, for both groups, the average delay between the sessions was longer than planned because of the restrictions caused by the global COVID-19 pandemic. After performing EPELI (either HMD or FSD) in the second session, the children were asked to evaluate, which version was more realistic, preferable, and easier to play, with response alternatives HMD/FSD/"I don't know". The parent or primary caregiver filled in the BRIEF questionnaire during both sessions, and the GEC raw score was used as the dependent variable.

The HMD-EPELI session included HMD-EPELI followed by the same oral questionnaires than in Study I and the WISC-IV Matrix reasoning and Similarities subtasks (Weschler, 2003). The participants who had already taken part in Study I were designated to the HMD-EPELI first group and did not perform that session again.

The FSD-EPELI session began with FSD-EPELI, which was followed by the Simulator Sickness Questionnaire (Kennedy et al., 1993) and the shortened version of the Presence Questionnaire 3.0 (Witmer et al., 2005; see also Supplementary Methods of Study I), which were read aloud and filled in by the experimenter for the lab group and by the parent for the home group. The group that performed FSD-EPELI at home also filled in a hardware questionnaire (see Supplementary Table 2 of Study IV).

3.4 Statistical analyses

Statistical analyses and data visualizations were done with *R* version 4.0.3 (Studies I, III, and IV) and version 4.2.0 (Study II; R Core Team, 2020). In addition to the base *R* packages, several other packages were employed. In Studies I, III, and IV, these included psych (Revelle 2020), tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), and ggplot2 (Wickham, 2016). In Study I, additional packages also included BayesFactor (Morey & Rouder, 2015), pROC (Robin et al., 2011), MatchIt (Ho et al., 2011), and rstatix (Kassambara, 2020). In Study II, additional packages were rcompanion (Mangiafico, 2023), and lmer (Bates et al., 2015). In Studies III and IV, data.table (Dowle & Srinivasan, 2021), stringr (Wickham, 2019), stringi (Gagolewski, 2020), dplyr (Wickham et al., 2021), gridExtra (Auguie, 2017), and patchwork (Pedersen, 2020) were also used. Finally, packages effectsize (Ben-Shachar et al., 2020) and ppcor (Kim, 2015) were employed in Study IV.

3.4.1 Study I

A small part (2.4 %) of the task performance data was missing because of technical failures and as one participant in the ADHD group refused to perform three tasks in the second measurement session. Before statistical analyses, first, the data of two participants (one from the control group and one from the ADHD group) was removed in the Cruiser task as they were purposefully colliding into other cars instead of avoiding them and said that this was more fun than the task they had been given. Second, participants who were not able to recall the PM task instructions after the HEXE, Cruiser or Clock task were excluded from the analyses of that task. Third, participants were removed from the analyses of the CPT, F&C, or HEXE task if they had performed near chance level (60 % or less on total correct answers). Fourth, all univariate outliers (± 3 standard deviations from the group mean) in the dependent variables were removed. Fifth, the data were controlled for possible multivariate outliers (Mahalanobis distance $\chi 2$ using alpha level $p < .001$) but none was found. The total amount of excluded data was 9.0 %.

From EPELI, five measures were used, namely Total score, Task efficacy, Navigation efficacy, Controller motion, and Total actions (see section 3.2.4). The background variables were tested for possible group differences with t-tests and Fisher's exact test.

The predictive validity of the EPELI measures for detecting group differences between the ADHD and control group were tested using two-way *analysis of variance* (ANOVA) with two independent variables, group (ADHD/control) and distractors (on/off). An additional ANOVA with scenario playing order (A/B) was the third independent variable was conducted, but as the results were very similar to those in the two-way ANOVA, they are not reported. All the assumptions of ANOVA were met, as evaluated through visual inspection. To examine the effects of scenario-to-scenario task progression on EPELI measures, general *linear mixed models* (LMMs) were used. In all LMMs, the random intercept model was the best fitting error covariance structure for all EPELI measures based on *the Bayesian Information Criterion* (BIC). The scenario-to-scenario variability of each EPELI measure and the dependent measures of conventional neuropsychological tasks were tested for group effects using t-tests. Bayes factors were calculated for all tests mentioned above.

The discriminant validity of each EPELI measure to classify individuals to the ADHD and control groups was evaluated by calculating the AUC from the ROC curve. A cutoff point with the highest percentage of correctly classified cases was determined with Youden's index (Youden, 1950), and the sensitivity and specificity at this cutoff was reported. The multivariate classification capacity of both EPELI and the CPT was evaluated by creating a composite measure from each task using logistic regression as group (ADHD/control) as the dependent variable and five measures from the corresponding task as the independent variables, and then evaluating the discriminant validity of this composite measure in the same way than for the individual variables. For EPELI, these five measures were Total score, Task efficacy, Navigation Efficacy, Controller motion, and Total actions, and for the CPT, they were omission errors, commission errors, reaction time variability, mean reaction time, and the number of correct responses.

The concurrent validity of the EPELI measures was examined by calculating Pearson's correlation coefficients between the EPELI measures and parent questionnaires (the BRIEF and ADHD-RS) over all participants, that is, the ADHD and control groups were combined for this analysis. As secondary analyses, correlations between the EPELI measures and conventional neuropsychological tasks and the parent questionnaires and conventional neuropsychological tasks were also calculated.

3.4.2 Study II

To control for possible differences between the ADHD and control groups in the background variables, t-tests and Mann-Whitney U-tests were used for age, average parental income, WISC-IV Matrix Reasoning, WISC-IV Similarities, the ADHD-RS total score, the CBCL internalizing and externalizing symptoms total score, the BRIEF total score, and the EQELI total score. For gender, Fisher's Exact test was used. All *p* values were adjusted for multiple comparisons with *False Discovery Rate correction* (FDR; Benjamini & Hochberg, 1995).

The differences between the ADHD and control group were tested with t-tests and Mann-Whitney U-tests in EPELI measures (Total score, Task efficacy, Navigation efficacy, Controller Motion, and Total actions), three eye movement measures (Fixation duration, Saccade duration, and Saccade amplitude) for each group of EPELI objects separately (attractive and relevant, attractive and irrelevant, non-attractive and relevant, non-attractive and irrelevant objects), three eye movement measures for two EPELI conditions (distracted vs. non-distracted), *Normalized Scanpath Saliency* (NSS), and Shoot the Target behavioral measures (Total score, Head rotation, and Head rotation speed).

To study classification accuracy, four *support vector machine* (SVM) classifiers were used. SVM1 was trained on the five EPELI measures (Total score, Task efficacy, Navigation efficacy, Controller motion, and Total actions), SVM2 on three eye movement metrics (Fixation duration, Saccade duration, and Saccade amplitude) averaged on each EPELI task scenario, and SVM3 on the same eye movement metrics using data from the Shoot the Target task. Additional SVM was trained on the same eye movement metrics using the instruction phase of each EPELI task scenario. The SVMs were trained in one cross-validation loop with automatic hyper-parameter search and evaluated in a separate loop with 10-fold cross-validation with the sklearn package version 1.0.02 in Python. The AUC on validation sample is reported. The data was first divided into ten folds for cross-validation samples, then scaled, decomposed with principal component analysis, and then used as input for each SVM classifier. The number of principal component analysis components, SVM kernel, C, and gamma parameters were select automatically using grid search. Differences in classifiers' performance were statistically tested with 30 iteration bootstrap cross-validation (see Fu et al., 2005).

3.4.2.1 Eye movement data processing

Two participants were excluded from the EPELI eye movement analysis because of high rate of invalid samples (30 % and 80 %). After the exclusion, there were no differences in the number of lost samples between the ADHD and control groups after, and the overall percentage of lost data was low (75 percentile, 5 %;

maximum, 16 %). In the Shoot the Target data, the percentage of invalid gaze samples was small in both groups (75 percentile, 1 %; maximum 10 %). However, the ADHD group had a higher percentage of lost data samples (Z = 3.00, p = .002). Therefore, the percentage of invalid gaze samples was included in the reported mixed models as a random effect.

Saccades and fixations were derived from raw gaze data using a modification of Engbert and Kliegl's algorithm (Engbert & Kliegl, 2003) for detecting eye movement events with unconstrained head movements (Larsson et al., 2016). Head movement compensation was performed by mapping raw gaze positions to world space coordinates. The eye movement features comprise saccade duration, saccade amplitude, and fixation duration. Peak velocity was not included to the analysis because of low sample rate of the eye tracker. Smooth pursuit detection was not included, since there were only few objects moving at a suitable speed in EPELI.

Saccades shorter than 22 ms (i.e., 1/sampling rate * 2), and with amplitude under 0.75 (following the suggestion by Larsson et al., 2016) were excluded from the analysis. Moreover, saccades longer than 300 ms were considered as noise and excluded, and the maximum saccade acceleration accepted was 36 000 degrees per second (see Larsson et al., 2016). Similarly, fixations shorter than 100 ms and longer than 1000 ms were removed.

From EPELI, only eye movements during the execution phase of each scenario were included in the analysis if not mentioned otherwise. Walls, the floor, and teleport waypoints were not considered as objects and fixations on them were not included in the analyses regarding relevance of the objects. They were also removed when calculating saccade proportions within the same object (for the analysis comparing ambient versus focal mode). However, an additional analysis was done for non-object related eye movements, in which fixations on the floor and walls where included. This was done to clarify the role of object interactions in group differences (see Study II Supplementary Results).

For saliency analysis, the participants view in the HMD was broadcasted to a tablet and recorded. The recording was cut into segments corresponding to each task based on time stamps. Each segment was processed using MT_TOOLS toolbox for creating saliency maps based on color, intensity, orientation, flickering, and motion (Bordier et al., 2013). For saliency index, NSS was used (Peters et al., 2005).

In the analysis of ambient versus focal mode, the execution phase of each EPELI scenario was divided into one second bins. For each bin, Number of Fixations, mean Fixation Duration, mean Saccade Amplitude and proportion of Saccades within the same object were calculated as features reflecting processing mode. The group differences in switches between the modes were examined with

the interaction effect between Group and Time on the selected features (See Study II Supplementary Methods).

3.4.3 Study III

Before the statistical analyses, the samples derived from Study I and II were tested for differences in age, gender, average parental income, and average parental education using t-tests and Fisher's Exact tests. As no differences were found (see Supplementary Appendix A in Study III), the samples were combined for all statistical analyses. The data were complete, except for the Instruction recall task, which one participant found oppressive and did not finish. In Study III, all eight EPELI measures (Total score, Task efficacy, Navigation efficacy, Controller motion, Total actions, TBPM score, Clock checks, and EBPM score) were analyzed (see section 3.2.4).

In the outlier analyses, the EPELI measures were checked for univariate outliers (± 3 standard deviations from the group mean). Three participants with outliers were excluded from the final sample, as all analyses included EPELI (final N = 77). The BRIEF measures (GEC, BRI, MI) were also checked for univariate outliers using same criterion, which resulted one more participant being removed from the analyses that included the BRIEF. The data were also controlled for multivariate outliers (Mahalanobis distance $\chi 2$ using alpha level $p < .001$) but none was found.

The internal consistency of each EPELI measure across scenarios was examined with the reliability measure of Cronbach's alpha using functions alpha and alpha.ci from psych package. Following an often-used interpretation, reliabilities of 0.7 and above were considered acceptable (see Nunnally & Bernstein, 1994).

The effects of background factors (age, gender, gaming background, familiarity with the tasks, and the HMD type) on each EPELI measure were tested with general linear models using function lm from stats package. The only collinearity between the independent variables (background factors) was that more boys than girls played regularly (for boys: 45 playing regularly and 1 not; girls: 22 playing regularly and 9 not; Fisher's Exact test $p < .001$). For each EPELI measure, the best fitting model was determined by three methods (forward, backward, and combination) using function step in package stats (for description of the methods, see Statistical analyses in Study III). All three selection methods resulted in the same models except for the measure of Clock checks. For Clock checks, the model acquired when starting from the null model was chosen as it had lower *Akaike information criterion* (AIC) than the model that resulted from starting from the full model. In the Results, the best fitting models are reported (for the full models, see Supplementary Appendix C in Study III). For all EPELI measures, the same

independent variables yielded statistically significant effects in both the best fitting and full model. To probe the effects of child's capacity for encoding verbal instructions on EPELI performance, general linear models with each EPELI measure at a time as the dependent variable and the Instruction recall task raw score as the independent variable were fitted to the data. The associations between the EPELI measures and parent-reported executive function difficulties (the BRIEF measures of GEC, BRI, MI) were quantified with Pearson's correlation coefficients. To study these relationships in greater detail, general linear models with each BRIEF measure at a time as the dependent variable and the EPELI measures as independent variables were fitted to the data, and the best fitting models were determined using the three selection methods (forward, backward, and combination) described above. Here, all three methods resulted in the selection of the same independent variables.

3.4.4 Study IV

Before statistical analyses, the data were inspected for missing values, data handling errors, and possible outliers. In the home group, the questionnaires to be answered after FSD-EPELI were missing from six participants and one parent had not answered the BRIEF questionnaire. Univariate outliers in EPELI, the BRIEF, and Presence questionnaire were first identified visually and confirmed numerically (± 3 SDs from the group mean). For FSD-EPELI, this was performed separately for the lab and home groups. Three HMD-EPELI gameplays, two FSD-EPELI gameplays, and two BRIEF questionnaires were excluded from the analyses because of outliers in at least one measure. The total amount of excluded data was 3.2 %. As in other studies, no multivariate outliers were found. The average administration time was equal between the EPELI versions, 27.5 minutes for the FSD and 27.8 minutes for the HMD version ($t(122) = -0.59$, $p = .55$).

Similarities and differences in task performance between the FSD and HMD version and the first and second session were examined with LMMs with each EPELI measure at a time as the dependent variable, EPELI version (FSD/HMD) and session (first/second) as fixed factors, and participant as a random factor. As an exception, the measure of Controller motion was not included in this analysis, as it measures somewhat different aspects in the two versions: in the FSD version the view is rotated using the controller (i.e., mouse/trackpad), while in the HMD this is done with head movements. In the models with Total actions and Clock checks as the dependent variable, the error terms were not normally distributed. For these measures, additional generalized LMMs with Poisson distribution were fitted. As these models yielded very similar results, only the general LMMs are reported. Cohen's d was used as the measure of effect size and interpreted as suggested by Cohen (1988) as small (> 0.20), medium (> 0.50), or large (> 0.80).

For the LMMs, the lmer function from the lme4 package was used, while Cohen's *d*s were obtained using the t_do_d function from package effectsize.

Similarities and differences in subjective experience between the FSD and HMD versions were studied as follows. LMMs with each Presence questionnaire at a time as the dependent variable, EPELI version and session as fixed factors and participant as a random factors were calculated. As the error term distributions in the models of questions 5, 6, 7, 8, and 12 were not normally distributed, the main effects of version and time were additionally tested with Wilcoxon signed-rank tests with continuity correction for these questions. These analyses yielded similar results as the LMMs and are therefore not shown. The three questions with head-to-head comparison of the two versions were tested with the exact binomial test.

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

The associations between the EPELI efficacy measures (Task and Navigation efficacy) and BRIEF measures were evaluated using bivariate correlations. As all distributions were near to normal based on visual inspection, Pearson's correlation coefficients were used. The correlations were calculated for both EPELI versions (HMD/FSD) and the corresponding BRIEF questionnaire, as well as for both EPELI sessions (first/second) and the corresponding BRIEF questionnaire.

To allow comparison with earlier literature, the inter-version correlations and test-retest stabilities of EPELI measures were first evaluated with bivariate correlation coefficients. To account not only for the within-subject change but also for the differences in the group means between the versions, *intraclass correlations* (ICCs) were also obtained using single-rating, absolute agreement, two-way random effect models (ICC 2,1 in Shrout & Fleiss, 1979) using function ICC from the psych package. To inspect the effect of one factor (version or session) while controlling for the other but without accounting for the within-subject variation, partial correlations were calculated for both inter-version and test-retest correlations with the other factor as a covariate. The partial correlations were obtained with the pcor function from package ppcor and were considered as the primary correlation measures. Based on visual inspection, all distributions in both EPELI versions were normally distributed, except those of Total actions, which were strongly skewed to the right. To evaluate if this skewness affected the results, these distributions were normalized with logarithmic transformations, and the inter-version and test-retest correlations were also calculated for the transformed variables. These results were almost identical (i.e., within ± 0.01

units) with those of the untransformed variables, and therefore only the results with the original variables are reported.

4 Results

4.1 Background characteristics

Background characteristics of the final samples in Studies I and II are shown in Table 2. In Study I, the ADHD and control groups were matching on age, handedness, gender, parental income, parental education, verbal reasoning abilities, and perceptual reasoning abilities. In Study II, the groups did not differ in terms of age or gender, but participants with ADHD had lower parental income and performed worse in verbal and perceptual reasoning tasks from the WISC-IV. In both studies, parent rated more inattention and hyperactivity-impulsivity symptoms (the ADHD-RS) and executive function problems (the BRIEF) for the children with ADHD than their controls. There were also more parent-reported psychiatric internalizing and externalizing symptoms (CBCL) and difficulties in real-life situations (EQELI) in the ADHD group. In neither study did the two groups differ in gaming experience, perceived familiarity of the tasks, or overall sense of presence. Both groups were able to reliably name the objects in EPELI (object naming task) and reported very few cybersickness symptoms.

Table 3 shows background characteristics of the final samples in Studies III and IV. As the sample of Study III was derived from the control groups of Studies I and II, the reported background characteristics are very close to those in Table 2. The sample of Study IV was also very similar, as 36 participants had earlier taken part in Study I as control participants and the inclusion criteria in Study IV were the same as those for the control participants in Study I. In both Study III and IV, the average parental income and education level was slightly higher than in the Finnish population aged 30 to 44 years[2]. In Study III, the verbal reasoning abilities were on average slightly higher ($t(77) = 4.41$, $p < .001$) and perceptual reasoning abilities on the same level ($t(77) = .061$, $p = .550$) than in the normative data reported in the Finnish WISC-IV test manual (Wechsler, 2003). Most children in Study II reported that they regularly played video, computer, or smartphone games.

[2] According to Official Statistics of Finland (2022a and 2022b), the average income before tax was 3406 €/m and the average education 2.3 (when converted to the scale used in the study) for adults aged 30 to 44 years.

Table 2 Background characteristics of the final samples in Studies I and II.

	Study I				Study II			
Variable	ADHD group (n = 38)	control group (n = 38)	Test statistic	*p*	ADHD group (n = 37)	control group (n = 36)	Test statistic	*p*
Age (years)	10.3 (1.1)	10.8 (1.1)	*t*(74) = -1.70	.093	10.3 (1.1)	10.8 (1.2)	*t*(69) = 1.62	.110
Handedness (right / left / ambidextrous)	32 / 5 / 1	36 / 1 / 1	Fisher's Exact	.200	-	-	-	-
Gender (boys / girls)	33 / 5	30 / 8	Fisher's Exact	.544	29 / 8	21 / 15	Fisher's Exact	.081
Parental income*	3.7 (1.0)	4.0 (1.0)	*t*(74) = -1.15	.253	**3.7 (1.1)**	**4.6 (0.7)**	***t*(66) = 4.34**	**< .001**
Parental education**	2.4 (0.6)	2.7 (0.5)	*t*(74) = -1.88	.063	-	-	-	-
WISC-IV Similarities	10.8 (2.4)	10.8 (3.1)	*t*(74) < 0.01	.999	**9.2 (3.4)**	**11.2 (2.4)**	***t*(63) = 2.80**	**.017**
WISC-IV Matrix reasoning	8.9 (2.7)	9.0 (3.3)	*t*(74) = -0.15	.879	**8.8 (4.0)**	**10.9 (3.5)**	***t*(69) = 2.37**	**.030**
ADHD-RS (0–56)	**31.5 (9.4)**	**7.9 (6.2)**	***t*(74) = 12.94**	**< .001**	**31.9 (8.3)**	**6.2 (6.5)**	***t*(66) = 14.63**	**< .001**
BRIEF GEC (0–144)	**87.7 (20.2)**	**31.5 (16.0)**	***t*(74) = 13.46**	**< .001**	**123.6 (25.0)**	**48.3 (24.1)**	***t*(69) = 12.93**	**< .001**
CBCL (0–126)	**54.6 (25.4)**	**17.9 (12.9)**	***t*(74) = 7.95**	**< .001**	**56.4 (23.0)**	**16.2 (14.4)**	***t*(59) = 8.85**	**< .001**
EQELI (0–60)	**41.3 (17.0)**	**12.8 (11.3)**	***t*(74) = 8.61**	**< .001**	**46.8 (14.8)**	**16.3 (12.7)**	***t*(68) = 9.35**	**< .001**
Gaming experience***	0.1 (1.9)	-0.1 (2.2)	*t*(74) = 0.37	.706	-	-	-	-
Familiarity of the tasks	5.1 (1.3)	4.9 (0.9)	*t*(74) = 0.72	.469	4.8 (1.7)	5.0 (1.1)	*t*(61) = 0.57	.760
Object naming task	19.5 (1.0)	19.6 (1.1)	*t*(74) = -0.45	.654	19.0 (0.8)	19.4 (0.7)	*t*(70) = 2.18	.130
Simulator sickness	0.6 (1.0)	0.9 (1.0)	*t*(74) = -1.11	.269	0.65 (1.0)	0.7 (1.1)	*t*(70) = 0.073	.940
Presence questionnaire	61.9 (11.3)	64.9 (8.5)	*t*(74) = -1.33	.187	70.4 (15.6)	72.3 (10.5)	*t*(63) = 0.59	.760

Note. The effects that are significant at $p < .05$ are written in bold. The numbers in parentheses indicate group standard deviation. * Before tax per adult: 1 = less than 1500 €/m, 2 = 1500–2200 €/m, 3 = 2200–3000 €/m, 4 = 3000–4000 €/m, 5 = over 4000 €/m. ** 1 = Comprehensive school, 2 = High school / Vocational school, 3 = University degree or equivalent. *** Sum of normalized scores of three questions 1) “How many days per week you play computer, console, or cell phone games?”, 2) “How long is your average playing session?", 3) "How many years have you played regularly?"

Table 3 Background characteristics of the final samples in Studies III and IV.

Variable	Study III (n = 77)	Study IV (n = 72)
Age (years)	10.8 (1.1)	11.0 (1.0)
Handedness (right / left)	71 / 6	67 / 5
Gender (boys / girls)	45 / 31	43 / 29
Parental income*	4.3 (0.83)	4.45 (0.79)
Parental education**	2.8 (0.41)	2.89 (0.30)
WISC-IV Similarities	11.3 (2.70)	-
WISC-IV Matrix reasoning	10.2 (3.38)	-
BRIEF GEC (0–144)	101.1 (15.77)	-
BRIEF BRI (0–56)	35.2 (5.48)	-
BRIEF MI (0–44)	65.4 (11.64)	-
Gaming background (regular / not)	67 / 10	-

Note. The numbers in parentheses indicate group standard deviation. * Before tax per adult: 1 = less than 1500 €/m, 2 = 1500–2200 €/m, 3 = 2200–3000 €/m, 4 = 3000–4000 €/m, 5 = over 4000 €/m. ** 1 = Comprehensive school, 2 = High school / Vocational school, 3 = University degree or equivalent.

4.2 Study I

4.2.1 Predictive and discriminant validity of the EPELI measures

Predictive validity. Figure 4 shows the distributions of the EPELI measures, and the effects of group and distractions on the same measures are presented in Table 4. For Total score, there were main effects of group and distractions, with the control group attaining higher scores than the ADHD group and the non-distracted scenarios resulting in higher scores than the distracted ones. Task efficacy and Navigation efficacy also had main effects of group and distractions with the control group being more efficient than the ADHD group and efficacy being higher in the non-distracted than distracted scenarios. Controller motion was also affected by main effects of group and distraction, as the ADHD group displayed more movement than the control group and the distracted scenarios yielding more motion than the non-distracted scenarios. For Total actions, there was a main effect of group, as the participants in the ADHD group performed more actions than those in the control group.

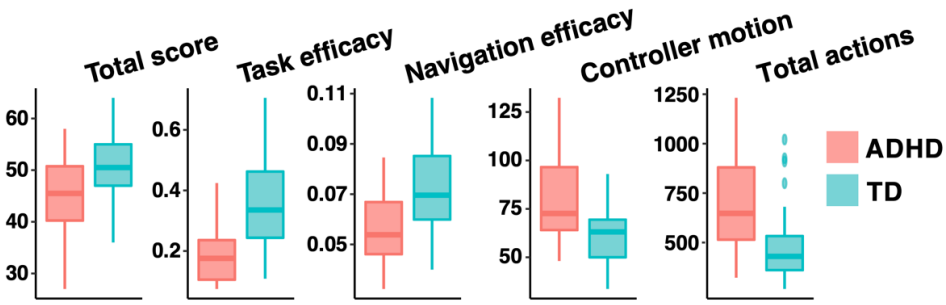


Figure 4 Boxplots showing the distributions of the EPELI measures per group (ADHD or typically developing [TD] controls). The vertical lines indicate the median. The dots in Total actions are outliers (± 1.5 * interquartile range from the mean). Adapted with permission from Study I.

Table 4 Analysis of variance for the effects of Group, Distractions, and Group x Distraction interaction for the EPELI measures.

Dependent variable	Effect	$F(1, 74)$	p	η^2_G	1 / BF
Total score	**Group**	**12.31**	**< .001**	**.12**	**39.03**
	Distractions	**26.97**	**< .001**	**.07**	**8550.36**
	Group x Distractions	2.50	.118	.01	0.60
Task efficacy	**Group**	**35.64**	**< .001**	**.28**	**145317.90**
	Distractions	**16.31**	**< .001**	**.04**	**167.64**
	Group x Distractions	0.113	.738	< .01	0.25
Navigation efficacy	**Group**	**16.67**	**< .001**	**.16**	**205.41**
	Distractions	**31.13**	**< .001**	**.07**	**34542.69**
	Group x Distractions	1.51	.223	< .01	0.45
Controller motion	**Group**	**16.79**	**< .001**	**.16**	**219.42**
	Distractions	**6.00**	**.017**	**.01**	**2.58**
	Group x Distractions	1.76	.188	< .01	0.51
Total actions	**Group**	**19.75**	**< .001**	**.19**	**579.12**
	Distractions	0.30	.585	< .01	0.21
	Group x Distractions	0.04	.840	< .01	0.24

Note. The effects that are significant at $p < .05$ are written in bold. η^2_G, effect size partial eta squared. BF, bayes factor.

In scenario-to-scenario progression (see Supplementary Figure 1 in Study I), there was a main effect of time on Task efficacy ($t[910] = -5.61$, $p < .001$), Navigation efficacy ($t[910] = -4.19$, $p < .001$), Controller motion ($t[910] = 3.92$, $p < .001$), and Total actions ($t[910] = 7.27$, $p < .001$) with the efficacy declining but the amount of motion and actions increasing as a function of time. These LMMs also showed main effects of group in all EPELI measures similar to those reported

in Table 4. Furthermore, an interaction effect between time and group was found on three measures. On Task efficacy (t[910] = 2.21, p < .027) and Navigation efficacy (t[910] = 2.43, p = .015) the ADHD group displayed stronger decline and on Total actions (t[910] = -3.05, p < .002) a stronger increase over time than the control group. In the scenario-to-scenario variability, a group effect was found in Task efficacy (t[74] = -3.67, p < .001), Controller motion (t[74] = 4.10, p < .001), and Total actions (t[74] = 3.53, p < .001), as the ADHD group displayed less variability in Task efficacy but more in the two other measures than the control group. As Task efficacy is the percentage of relevant actions of total actions (excluding moving actions), separate analyses for the variabilities of its constituent measures were conducted to better interpret the group effect in the variability of Task efficacy. There was no group effect in the number of relevant actions, but total actions excluding moving actions had a group effect with the ADHD group demonstrating more variability than the control group (t[74] = 3.53, p < .001). Therefore, the group difference in Task efficacy results from more variability in total actions in the ADHD group.

Discriminant validity. Table 5 shows the AUCs and cutoff values for the EPELI measures and logistic regression composite variable based on all five EPELI measures. Of the single EPELI measures, Task efficacy yielded the highest AUC (.83). The composite logistic regression analysis measure had a slightly higher AUC (.88), but based on the confidence intervals, the difference to the AUC of Task efficacy was not significant. The ROC curve of the composite logistic regression analysis measure is shown in Figure 5.

Table 5 Area under the curves (AUCs) from ROC analyses, 95 % confidence intervals and optimal cutoffs for each EPELI measure and logistic regression analysis composite utilizing all five EPELI measures at the same time.

			Optimal cutoff*		
Variable	AUC	95 % CI	Threshold	Sensitivity	Specificity
Total score	.70	.59–.82	46.5	76 %	55 %
Task efficacy	.83	.74–.92	.29	66 %	89 %
Navigation efficacy	.75	.64–.86	.06	76 %	66 %
Controller motion	.73	.62–.85	68588.85	71 %	66 %
Actions	.78	.68–.89	463	61 %	89 %
Logistic regression analysis	.88	.80–.94	.431	79 %	87 %

Note. * Based on Youden's index. AUC, area under the curve. CI, confidence interval.

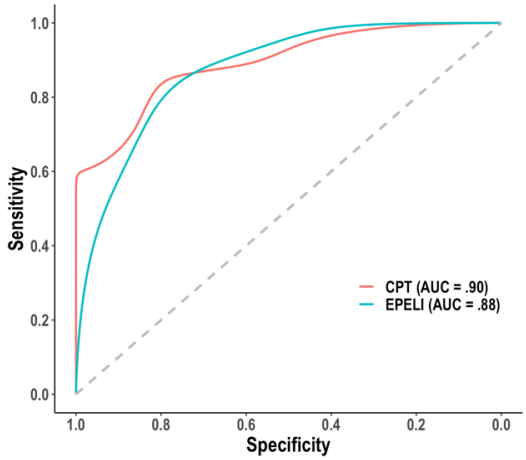


Figure 5 Receiver operating characteristic (ROC) curves of the logistic regression analysis measures from five EPELI and five CPT measures. Adapted with permission from Study I.

4.2.2 Group differences and discriminative ability of the conventional neuropsychological tasks

The group (ADHD/control) differences on the conventional neuropsychological tasks are shown in Table 6 and the distributions of the tasks with significant group differences in Figure 6. The control group had better performance than the ADHD group in the Digit span task but not in the Instruction recall task, where the material was similar to the instructions heard in EPELI. Regarding the CPT, the participants with ADHD made more omission and commission errors and had higher variability in reaction time than the controls. Moreover, the ADHD group showed longer mean reaction times in SRT and a higher switching cost in F&C. In the PM tasks, the ADHD group performed worse than the control group in the time-based PM task Cruiser, but no group effects were found in the event-based PM Clock task or the PM measures in the HEXE task (self-initiation and switching). As regards to ongoing task performance of the PM tasks, the ADHD participants made more mistakes than the controls both in the Cruiser (the number of crashes) and HEXE (ongoing errors) tasks, but there was no group difference in the number of correct ongoing task responses in HEXE. The control participants engaged in more active time monitoring in the Cruiser task than the participants with ADHD.

Table 6 Group effects in the conventional neuropsychological tasks.

Variable	Test statistic	*p*	1 / BF	Effect size*
Digit span	***t*(74)= -2.78**	**.007**	**6.14**	**.638**
Instruction recall task	*t*(74) = 1.04	.301	0.38	.239
CPT omissions	***t*(65) = -3.16**	**.002**	**14.80**	**-.786**
CPT commissions	***t*(65) = -2.91**	**.005**	**8.08**	**-.711**
CPT RT variability	***t*(65) = -5.54**	**< .001**	**21755.35**	**-1.38**
SRT mean RT	***t*(73) = -2.98**	**.004**	**9.62**	**-.684**
F&C switching cost	***t*(66) = -3.21**	**.002**	**16.64**	**-.765**
Cruiser PM accuracy	***t*(70) = 3.83**	**< .001**	**91.93**	**.913**
Cruiser monitoring	***t*(70) = 2.23**	**.029**	**1.98**	**.527**
Cruiser the number of crashes	***t*(70) = -2.48**	**.016**	**3.20**	**-.589**
Clock task PM accuracy	*t*(72) = 0.47	.637	0.25	.110
HEXE correct task responses	*t*(57) = -1.31	.195	0.54	-.335
HEXE ongoing errors	***t*(57) = -2.08**	**.028**	**1.57**	**-.544**
HEXE self-initiated PM task	Fisher's Exact test	.332	0.64	.094
HEXE switching PM task	Fisher's Exact test	.691	0.51	< .001

Note. The effects significant at the level of $p < .05$ are written in bold. PM, prospective memory. RT, reaction time. BF, bayes factor. * Cohen's *d* for continuous variables, Cramér's *V* for categorial variables.

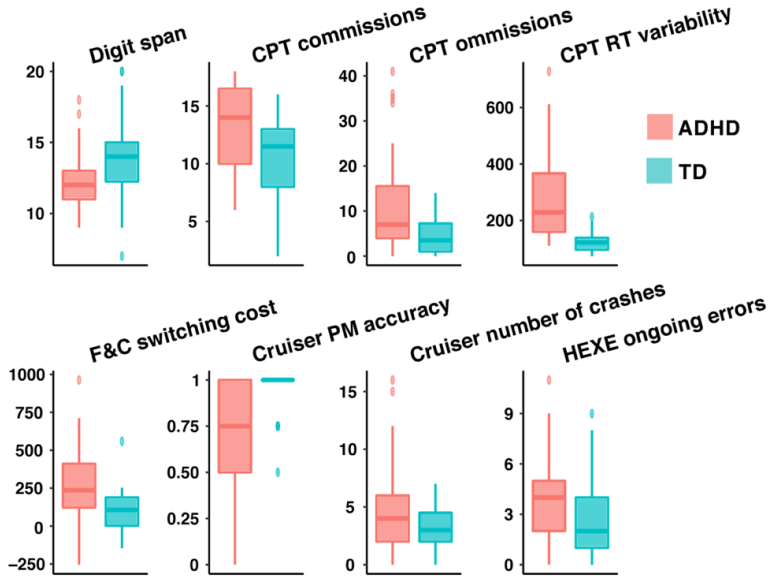


Figure 6 Boxplots showing the distributions of the conventional neuropsychological tests, which showed group differences, per group (ADHD/control). The vertical lines indicate the median. The dots are outliers (± 1.5 * interquartile range from the mean). Adapted with permission from Study I.

The AUCs and cutoff values with optimal discrimination ability for the conventional neuropsychological tasks and logistical regression analysis measure based on all five CPT measures are presented in Table 7. The highest AUC (.90) was that of the logistic regression measure in the CPT, but based on the confidence intervals, this was not statistically higher than the AUC for CPT reaction time variability (.85). When considering the AUCs of EPELI and the conventional neuropsychological tasks together, the highest were those yielded by the EPELI logistic regression analysis, EPELI Task efficacy, the CPT logistic regression analysis, and CPT reaction time variability, which did not differ from each other (all p values > .05). However, the AUC of EPELI Task efficacy was significantly higher than that of the conventional neuropsychological tasks, except Digit span, CPT reaction time variability, F&C switching cost, and HEXE ongoing errors.

Table 7 Area under the curves (AUCs), 95 % confidence intervals and optimal cutoffs from the ROC analyses for each conventional neuropsychological task and logistic regression analysis composite utilizing all five CPT measures at the same time.

			Optimal cutoff*		
Variable	AUC	95 % CI	Threshold	Sensitivity	Specificity
Digit span	.70	.58–.82	12.5	74 %	66 %
Instruction recall task	.47	.34–.60	25.5	39 %	66 %
CPT omission errors	.70	.57–.82	3.5	50 %	80 %
CPT commission errors	.70	.58–.82	13.5	78 %	51 %
CPT RT variability	.85	.76–.94	150.32	88 %	77 %
SRT mean RT	.67	.54–.79	508.12	58 %	73 %
F&C switching cost	.71	.58–.84	274.01	97 %	47 %
Cruiser PM accuracy	.70	.60–.80	0.88	86 %	51 %
Cruiser number of clock checks	.66	.53–.78	18.5	77 %	54 %
Cruiser number of crashes	.63	.50–.76	7.5	100 %	22 %
Clock task PM accuracy	.48	.36–.60	0.0	0 %	100 %
HEXE correct task responses	.55	.38–.72	41	82 %	44 %
HEXE ongoing errors	.67	.53–.81	2.5	62 %	72 %
CPT logistic regression analysis	.90	.82–.96	0.43	81 %	86 %

Note. * Based on Youden's index. RT, reaction time. AUC, area under the curve. CI, confidence interval.

4.2.3 Concurrent validity of the EPELI measures

Regarding concurrent validity, correlations of EPELI and the BRIEF and ADHD-RS questionnaires across all participants are shown in Table 8. All EPELI

measures correlated with both the BRIEF and ADHD-RS (absolute *r* values [.312–.574]). For EPELI Total score, Task efficacy, and Navigation efficacy, this association was negative, meaning that higher performance in these measures correlated with fewer executive function problems and lower ADHD symptom scores. For Controller motion and Total actions, the direction was the opposite.

Table 8 Correlations between the EPELI and conventional neuropsychological measures and the BRIEF and ADHD-RS questionnaires.

	BRIEF		ADHD-RS	
Variable	*r*	95 % CI	*r*	95 % CI
EPELI Total score	**-.356 ****	[-.539, -.142]	**-.312 ***	[-.502, -.093]
EPELI Task efficacy	**-.574 *****	[-.708, -.400]	**-.553 *****	[-.692, -.375]
EPELI Navigation efficacy	**-.466 *****	[-.626, -.269]	**-.453 *****	[-.615, -.253]
EPELI Controller movement	**.414 *****	[.208, .585]	**.430 *****	[.227, .598]
EPELI Total actions	**.457 *****	[.258, .619]	**.477 *****	[.282, .634]
Digit span	**-.304 ***	[-.496, -.084]	-.249	[-.449, -.024]
Instruction recall task	-.144	[-.358, .084]	-.057	[-.279, .171]
CPT omissions	**.368 ****	[.141, .559]	**.38 ****	[.153, .568]
CPT commissions	**.329 ***	[.096, .527]	**.353 ****	[.123, .546]
CPT variability	**.476 *****	[.266, .643]	**.469 *****	[.258, .638]
SRT mean RT	**.372 ****	[.159, .553]	**.335 ****	[.116, .522]
F&C switching cost	.219	[-.020, .435]	.175	[-.066, .397]
Cruiser PM accuracy	**-.329 ***	[-.521, -.106]	**-.288 ***	[-.487, -.061]
Cruiser monitoring	-.213	[-.424, .019]	-.18	[-.395, .054]
Cruiser number of crashes	.232	[.000, .440]	.146	[-.089, .365]
Clock task PM accuracy	-.042	[-.268, .189]	-.033	[-.260, .197]
HEXE correct task responses	.225	[-.033, .455]	.227	[-.031, .456]
HEXE ongoing errors	.257	[.001, .481]	.267	[.012, .490]
HEXE self-initiated PM task	.017	[-.238, .270]	.073	[-.184, .321]
HEXE switching PM task	-.064	[-.195, .315]	-.027	[-.281, .231]

Note. The correlations that are significant at the level of p < .05 are written in bold. * $p < .05$. ** $p < .01$. *** $p < .001$. FDR adjusted point estimates with unadjusted 95 % confidence intervals (CI).

4.2.4 Associations between the conventional neuropsychological tasks and parent-rated EF deficits and ADHD symptoms

The correlations between the conventional neuropsychological tasks and the BRIEF and ADHD-RS questionnaires across all participants are shown in Table 8 (absolute *r* values [.017–.476]). The CPT measures and SRT reaction time were positively correlated with both questionnaires, Digit negatively with the BRIEF,

and PM accuracy in the Cruiser negatively with both the BRIEF and ADHD-RS. The correlation between EPELI Task efficacy and the BRIEF was stronger than any of the correlations between the conventional neuropsychological tasks and the BRIEF, except for that between the CPT reaction time variability and the BRIEF, when the unadjusted confidence intervals were used as the criterion.

4.2.5 Associations between the EPELI measures and conventional neuropsychological tasks

The correlations between the EPELI measures and conventional neuropsychological tasks that yielded group differences (see Table 6) are displayed in Table 9. The commission errors in the CPT correlated positively with Controller motion and Total actions in EPELI. The omissions in the CPT showed a negative correlation with EPELI Task and Navigation efficacy and a positive correlation with EPELI Total actions. The reaction time variability in the CPT was negatively associated with EPELI efficacy measures and positively correlated with EPELI Controller motion and Total actions. As regards to the mean reaction time in SRT, negative correlations to EPELI Total score and efficacy measures were found. The switching cost in the F&C task correlated negatively with EPELI Total score. Regarding PM, the PM performance in the Cruiser task correlated positively with EPELI Total score, Task efficacy, and Navigation efficacy. The ongoing task performance in HEXE correlated negatively with EPELI Total score and efficacy measures but positively with EPELI Controller motion and Total actions.

Table 9 Correlations between the EPELI measures and conventional neuropsychological measures that yielded group differences.

	Digit span	CPT commissions	CPT omissions	CPT RT variability	SRT mean RT	F&C switching cost	Cruiser PM accuracy	Cruiser number of crashes	HEXE Ongoing errors
Total score	.244	-.151	-.181	-.252	**-.383 ****	**-.315 ***	**.452 *****	-.110	**-.354 ***
Task efficacy	.167	-.260	**-.346 ***	**-.407 ****	**-.278 ***	-.218	**.338 ***	-.135	**-.515 *****
Navigation efficacy	.211	-.254	**-.290 ***	**-.351 ***	**-.331 ***	-.220	**.358 ****	-.061	**-.425 ****
Controller motion	-.134	**.295 ***	.193	**.294 ***	.072	.115	-.192	-.001	**.566 *****
Total actions	-.171	**.312 ***	**.375 ****	**.413 ****	.220	.093	-.197	.071	**.581 *****

Note. The correlations that are significant at $p < .05$ are written in bold. * $p \leqq .05$. ** $p \leqq .01$. *** $p \leqq .001$. RT, reaction time.

The correlations between the EPELI measures and other conventional neuropsychological tasks (i.e., those not yielding group differences) can be in Supplementary Table 3 of Study I. The Instruction recall task correlated positively with EPELI Total score and efficacy measures but negatively with EPELI Controller motion and Total actions. Both WISC-IV reasoning tasks were positively associated with EPELI Total score and Similarities also with EPELI efficacy measures. Regarding PM, time monitoring in the Cruiser task positively correlated with EPELI Total score, and the total number of correct task responses in HEXE with EPELI Controller motion and Total actions.

4.3 Study II

4.3.1 Task performance

The distributions of the EPELI measures per group can be seen in Figure 7. The ADHD group had a lower Total score (Z = 2.31, p = .021), Task efficacy (Z = 3.10, p = .005), and Navigation Efficacy (Z = 2.73, p = .009), and a higher level of Controller motion (Z = 2.70, p = .009) and more Total actions (Z = 3.20, p = .005), compared to the control group. This finding replicated the group effects found in Study I (see Figure 4 and Table 4).

To study how well the EPELI measures can classify participants to the ADHD and control groups in these data, a SVM classifier was trained based on these five performance measures (SVM 1). This analysis yielded an AUC score of .77.

In Shoot the Target task, the ADHD group acquired lower Total score than the control group (t(54) = 3.63, p = .002). There were no group differences in Head rotation angle and Head rotating speed (see also Supplementary Results Figure S3 in Supplementary Material of Study II).

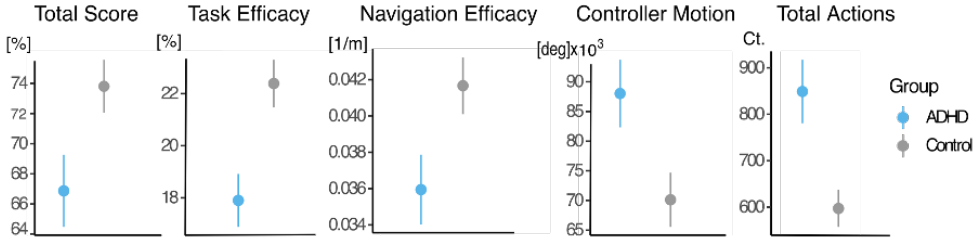


Figure 7 The distributions of EPELI performance measures per group (ADHD/control) in Study II. Error bars indicate the standard error of mean. Adapted permission from Study II.

4.3.2 Eye movement behaviors

The effects on fixations and saccades across the EPELI scenarios. LMMs were used to analyze Saccade duration, Saccade amplitude, and Fixation duration across the EPELI scenarios. The models included Group and Task scenario as fixed effects and Participant and Task scenario order (forward/reversed) as random effects (see Table 10). A possible distractor effect was accounted for by including Task scenario order as a random factor, as different scenarios were embedded with the distractions in the two task scenario orders. The LMMs revealed that the ADHD group had overall shorter Saccade durations and smaller Saccade amplitudes, and the effect was consistent across EPELI scenarios (see Figure 8). On average, Fixation durations were longer in the ADHD group than control group, but this effect was moderated by Task scenario (see Table 10 and Figure 8).

Table 10 Test statistics and linear mixed effect models for the eye movement features in EPELI.

Dependent variable	ADHD group Mean (SD)	Control group Mean (SD)	Effect	χ^2	df	φ^2	*p*
Saccade duration (ms)	57 (8.3)	67 (10.1)	**Task scenario**	**186.5**	**12**	**1.74**	**< .001**
			Group	**17.95**	**1**	**0.57**	**< .001**
			Task scenario x Group	13.20	12	0.49	.35
Saccade Amplitude (deg)	5.44 (1.4)	6.29 (1.7)	**Task scenario**	**243.5**	**12**	**2.01**	**< .001**
			Group	**9.99**	**1**	**0.44**	**.002**
			Task scenario x Group	18.94	12	0.56	.14
Fixation duration (ms)	317 (36)	309 (30)	**Task scenario**	**128.6**	**12**	**1.43**	**< .001**
			Group	**3.93**	**1**	**0.25**	**.047**
			Task scenario x Group	**25.47**	**12**	**0.64**	**.038**

Note. The effects that are significant at $p < .05$ are written in bold. φ^2, omega squared.

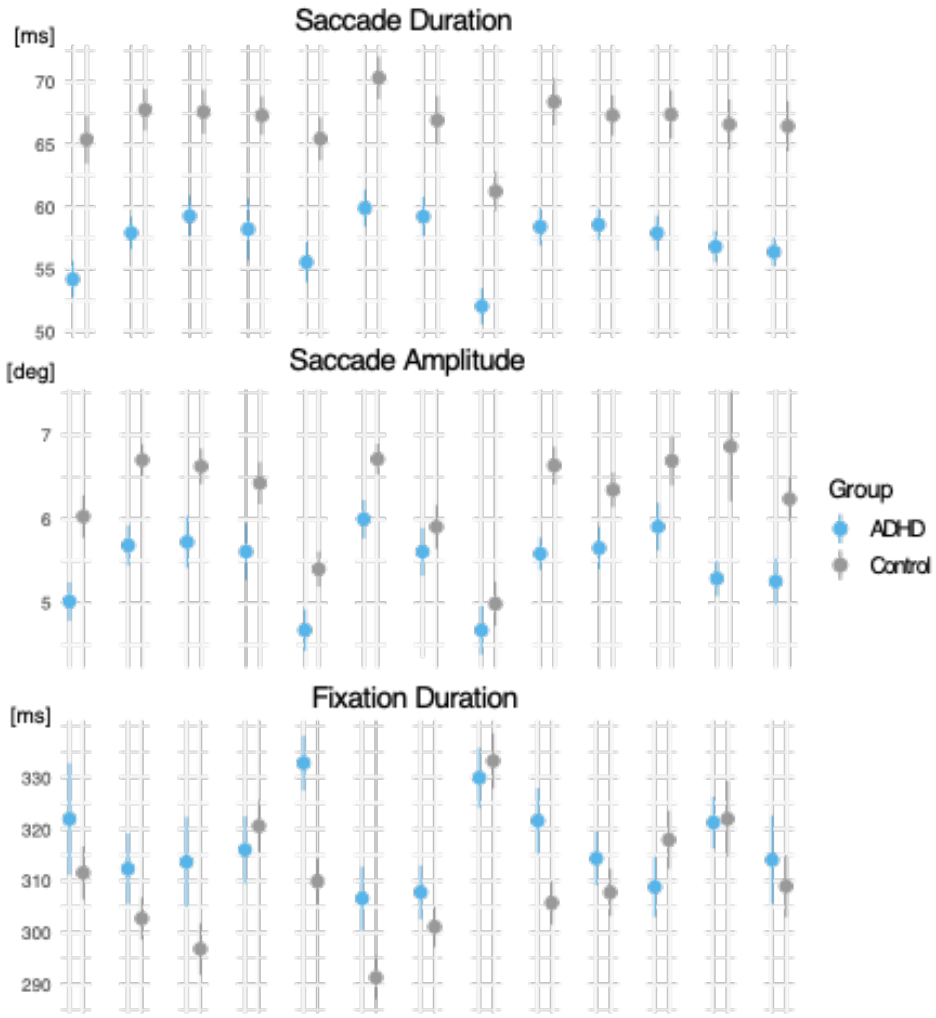


Figure 8 The effects of Task Scenario and Group on the eye movement features in EPELI per group. Each column pair represents one EPELI scenario. Error bars indicate the standard error of mean. Adapted permission from Study II.

To study the main research question related to the classification capacity of the eye movement features in EPELI, we trained an SVM classifier (SVM 2) on Saccade duration, Saccade amplitude, and Fixation duration data aggregated per task scenario. This classifier demonstrated an excellent AUC score of .92 after tenfold cross-validation (see Figure 9), which was higher than that of SVM 1 trained on the five EPELI performance measures ($t(58) = 8.40$, $p < .001$).

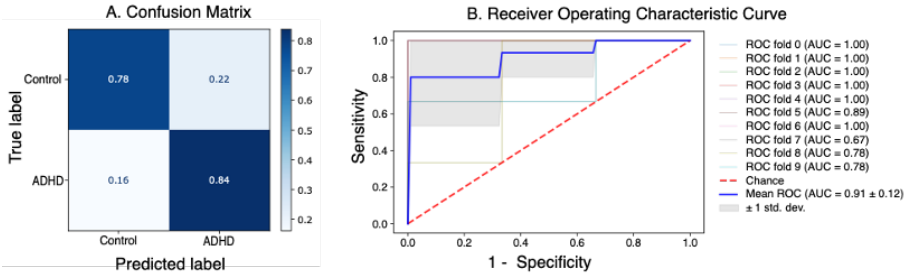


Figure 9 Evaluation of the SVM classifier trained on EPELI eye movement features (SVM 2). (A) Confusion matrix that shows the percentage of correctly classified and mislabeled individuals per group averaged on tenfold validation, (B) panel that shows the ROC curve and AUC scores fold and the average. Adapted with permission from Study II.

The effect of task performance on the eye movements in EPELI. To probe the role of task performance on the group effects in eye movement features, the LMMs testing the group effects were reran by regressing out the effect of Total score. The group differences remained significant in adjusted Saccade duration ($\chi 2(1) = 14.60$, $\varphi = 0.48$, $p < .001$) and Saccade amplitude ($\chi 2(1) = 5.90$, $\varphi = 0.31$, $p = .002$). The LMM with adjusted Fixation duration no longer showed a significant group effect ($\chi 2(1) = 2.44$, $\varphi = 0.20$, $p = .120$), but the interaction between Task scenario and Group remained significant ($\chi 2(12) = 25.45$, $\varphi = 0.64$, $p = .038$).

The effect of saliency on task performance in EPELI. The impact of saliency on participants' performance was examined with a LMM that included Total score as the dependent variable, NSS calculated per task scenario, Task scenario, and Group as fixed effects, and participant as a random effect. The results (see Table 11) showed a difference in difficulty between EPELI scenarios (the effect of Task scenario on Total score) and a group effect with the controls performing better than the children with ADHD. However, there was no effect of NSS nor an interaction effect of NSS and Group. Moreover, NSS did not differ between the groups, across all EPELI scenarios, or when the distracted and non-distracted scenarios were tested separately. On average NSS, was near to zero in both groups (0.21 ± 1.4 in the ADHD and 0.02 ± control group), suggesting that saliency had only a weak effect on gaze allocation.

Table 11 The linear mixed model of normalized scanpath saliency (NSS) on Total score.

Dependent variable	Effect	$\chi 2$	df	φ	p
Total score*	**Task scenario**	**124.89**	**12**	**1.41**	**< .001**
	NSS*	1.96	1	0.18	.150
	Group	**6.05**	**1**	**0.31**	**.014**
	NSS x Group	0.513	1	0.08	.510

Note. The effects that are significant at $p < .05$ are written in bold. φ^2, omega squared.

The effects of object relevance in EPELI. To further clarify which factors drive the observed group effects on eye movements, it was investigated how the relevance of objects in EPELI affected attention allocation (see Figure 10). Overall, Fixation duration was different between task-relevant and task-irrelevant objects (Z = 8.6, p < .001). However, there was no group differences between the task-irrelevant and task-relevant objects. The ADHD group had a tendency for longer Fixation durations to the task-irrelevant objects than the control group, but the effect was not significant (Z = 2.2, p = .060).

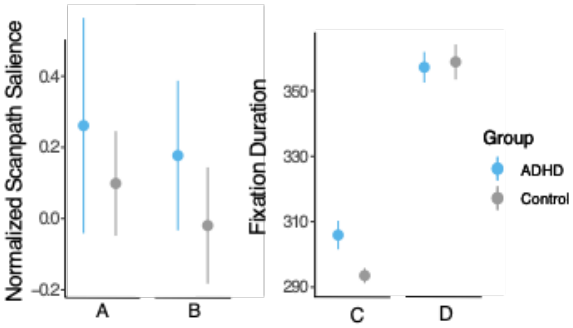


Figure 10 The left panel shows normalized scan path saliency (NSS) averaged by Condition. (A) Task scenarios without distractions, (B) Task scenarios with distractors. The right panel represents fixation durations to different types of objects. (C) Task-irrelevant objects, (D) Task-relevant objects. Adapted with permission from Study II.

The effects of ambient versus focal processing in EPELI. There was an effect of time on eye movement features, which suggests that the participants were indeed switching between ambient and focal processing during EPELI (see Figure 11). However, there were no group differences in switching between the ambient and focal modes. An additional analysis revealed that the group effects on the eye movement features where markedly different in the instruction and task execution phases of each scenario (see Supplementary Results of Study II).

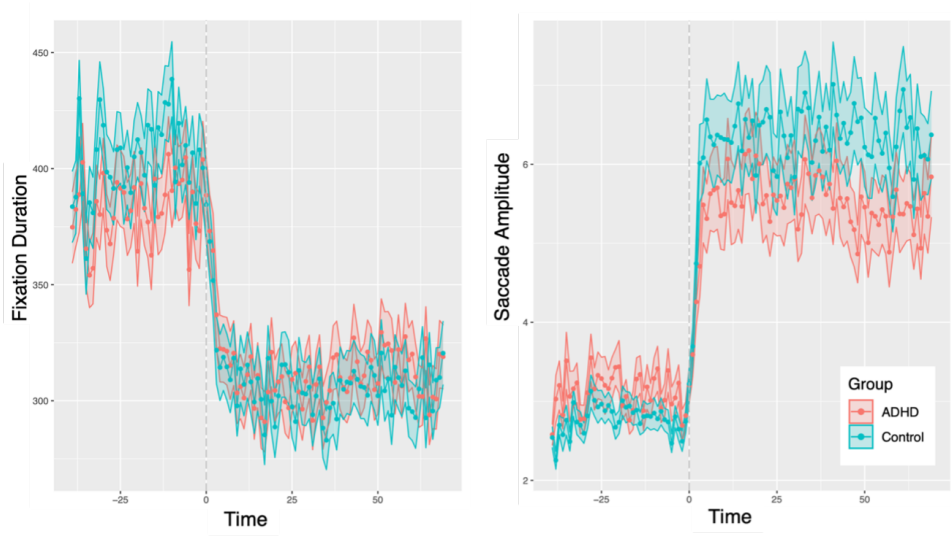


Figure 11 Eye movement characteristics over time. Zero time point is the time when the instruction phase ended and task execution phase began. (A) Fixation duration, (B) Saccade amplitude. For features tested in the analyses, see Figure 1 in Supplementary Material of Study II. Adapted with permission from Study II.

The effects of eye movement features in Shoot the Target. Like in EPELI, the LMMs in Shoot the Target task showed that the ADHD participants had shorter Saccade durations ($\chi 2(1) = 10.25$, $\varphi = 0.41$, $p = .002$) and Saccade amplitudes ($\chi 2(1) = 4.17$, $\varphi = 0.26$, $p = .041$) and longer Fixation durations ($\chi 2(1) = 12.04$, $\varphi = 0.44$, $p = .002$).

To accommodate for the difference in data loss between the groups (see section 3.4.2.1), an additional LMMs with the percentage of invalid gaze samples as a random factor was fitted to the data. In these models, the group difference in Fixation duration was not affected by the lost data (group effect: $\chi 2(1) = 8.43$, $p = .011$, percentage of invalid gaze samples effect: $\chi 2(1) = 0.02$, $p = .87$), but there was no significant group effect on Saccade duration (group effect: $\chi 2(1) = 3.04$, $p = .12$, percentage of invalid gaze samples effect: $\chi 2(1) = 6.58$, $p = .03$) or Saccade amplitude (group effect: $\chi 2(1) = 1.41$, $p = .23$, percentage of invalid gaze samples effect: $\chi 2(1) = 1.40$, $p = .36$).

To examine the classification capacity of eye movement features in Shoot the Target, another SVM classifier (SVM 3) was trained. This classifier yielded AUC score of .78, which was worse than that of SVM 2 based on the eye movement features in EPELI ($t(58) = 9.93$, $p < .001$).

4.4 Study III

4.4.1 Reliability of the EPELI measures

The internal consistency was acceptable (Cronbach's α >= .70) for all EPELI measures except TBPM and EBPM score (Table 12). The average consistency of the measures was .79 when TBPM and EBPM scores, which are included in Total score, were not also separately considered. When the average consistency was evaluated using all eight EPELI measures without considering the dependency between Total score and TBPM and EBPM scores, it was .71.

When shorter EPELI sections were evaluated by decreasing the number of scenarios stepwise from 13, the consistency remained acceptable down to five scenarios for Total actions, to seven scenarios for Task efficacy and Controller motion, and to 11 scenarios for Navigation efficacy and Clock checks. For Total score, it dropped under .70 in 12 scenarios (see Supplementary Appendix B in Study III). As regards to improving the internal consistency, dropping out the scenario least inconsistent with the others increased the internal consistency of Total score from .70 to .73.

Table 12 Internal consistency of the EPELI measures.

Measure	α (95 % CI) *
Total score	.70 [.59, .79]
Task efficacy	.83 [.77, .88]
Navigation efficacy	.74 [.65, .81]
Controller motion	.88 [.85, .92]
Total actions	.87 [.82, .91]
TBPM score	.59 [.45, .71]
Clock checks	.72 [.62, .80]
EBPM score	.33 [.13, .54]

Note. N = 77. * a Bootstrap 95 % Confidence Interval. TBPM, time-based prospective memory score. EBPM, event-based prospective memory score.

4.4.2 Associations between the EPELI measures and background factors

Table 13 shows the best fitting linear models with each EPELI measure as the dependent variable and background factors (gender, age, gaming background, task familiarity, HMD type) as the independent variables (for full models, see Supplementary Appendix C in Study III). Gender differences were found in five measures as girls obtained higher Total scores, higher Task and Navigation

efficacies, and performed fewer Total actions and Clock checks. Regarding age, older children acquired higher Total and TBPM scores, better Navigation efficacies and perform fewer actions. For the measures with both gender and age effects, additional models with interaction between these independent variables were fitted, but no interactions effects were present. Scatter plots of EPELI measures per age and gender are shown in Figure 12 (for descriptive statistics, see Supplementary Appendix I in Study III). For all EPELI measures the amount of variance explained was low (adjusted R^2 = .04–.20).

Table 13 The best fitting models with each EPELI measure as the dependent variable and age, gender, gaming background, familiarity of the tasks, and the head-mounted display (HMD) type as independent variables.

Dependent variable	Independent variable	Estimate (β)	SD	*t*	*p*	△AIC	R2	Adj. R^2
Total score	**(Intercept)**	**21.253**	**8.166**	**2.60**	**.011**	-3.77	.192	.159
	Gender	**4.524**	**1.615**	**2.80**	**.007**			
	Age (years)	**2.326**	**0.700**	**3.33**	**.001**			
	Gaming background	4.272	2.344	1.82	.072			
Task efficacy	(Intercept)	0.138	0.153	0.90	.371	-3.04	.185	.151
	Gender	**0.099**	**0.029**	**3.34**	**.001**			
	Age (years)	0.024	0.013	1.77	.082			
	HMD	-0.051	0.030	-1.74	.085			
Navigation efficacy	(Intercept)	0.010	0.021	0.46	.645	-2.97	.220	.199
	Gender	**0.014**	**0.004**	**3.35**	**.001**			
	Age (years)	**0.006**	**0.002**	**3.25**	**.002**			
Controller motion	**(Intercept)**	**92835**	**18483**	**5.02**	**<.001**	-5.33	.073	.048
	Gender	-7095	3580	-1.98	.051			
	Age (years)	-2664	1685	-1.58	.118			
Total actions	**(Intercept)**	**836.820**	**157.34**	**5.32**	**<.001**	-5.01	.137	.114
	Gender	**-85.100**	**30.480**	**-2.79**	**.007**			
	Age (years)	**-32.570**	**14.340**	**-2.27**	**.026**			
TBPM score	(Intercept)	-3.586	2.921	-1.23	.223	-2.16	.130	.107
	Age (years)	**0.787**	**0.266**	**2.95**	**.004**			
	Gender	1.038	0.566	1.83	.070			
Clock checks	**(Intercept)**	**34.696**	**1.919**	**18.08**	**<.001**	-3.02	.055	.042
	Gender	**-6.309**	**3.025**	**-2.09**	**.040**			
EBPM score	**(Intercept)**	**3.870**	**0.118**	**32.79**	**<.001**	-6.42	.039	.026
	Gender	0.324	0.186	1.74	.086			

Note. N = 77. The effects that are significant at $p < .05$ are written in bold. For gender, girl = 1 and boy = 0. For gaming background, regular = 1, not regular = 0. For HMD, Pico Neo 2 Eye = 1, Oculus GO = 0. △AIC, change in the Akaike Information Criterion from the full model.

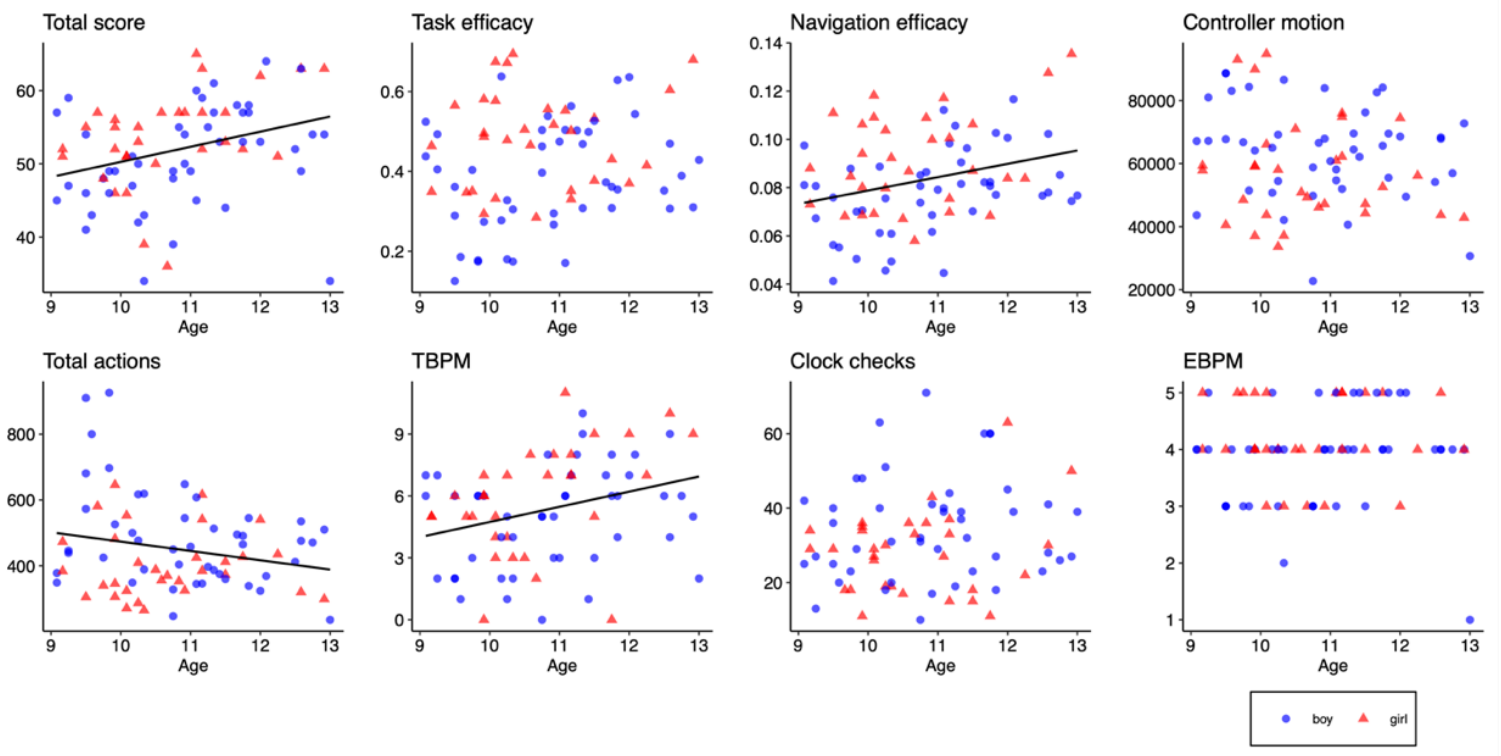


Figure 12 Scatter plots of the EPELI measures per age and gender. The regression line between an EPELI measure and age is shown on plots in which the correlation between the two was statistically significant. Adapted with permission from Study III.

To facilitate the interpretation of the associations presented above, the best fitting models were reanalyzed with Instruction recall task score as an additional independent variable (Supplementary Appendix D in Study III). In these models, instruction recall task performance had an effect on Total score, Task efficacy, Navigation efficacy, and EBPM score. Compared to the models presented above, the gender effect on Total score and the age effect on Total actions become non-significant, but in Controller motion a gender effect was found (boys > girls).

Even though gaming background, familiarity of the task contents, and HMD type showed no statistically significant effects on the EPELI measures (Table 13), there was a difference in the perceived hand controller quality as the Pico Neo 2 Eye group reported fewer problems than the Oculus GO group (see Supplementary Appendix G in Study III).

4.4.3 Associations between the EPELI measures and Instruction recall task performance

Higher performance in the Instruction recall task was associated with higher Total and EBPM score, better Task efficacy and Navigation Efficacy, and fewer Total actions (adjusted R^2 = .07–.18, see Table 14).

Table 14 The models with each EPELI measure as the dependent variable and the Instruction recall task as an independent variable.

Dependent variable	Independent variable	Estimate (β)	SD	*t*	*p*	R^2	Adj. R^2
Total score	(Intercept)	**42.645**	**2.352**	**18.13**	**<.001**	.194	.183
	Instruction recall task	**0.309**	**0.073**	**4.23**	**<.001**		
Task efficacy	(Intercept)	**0.245**	**0.049**	**4.95**	**<.001**	.162	.151
	Instruction recall task	**0.006**	**0.002**	**3.79**	**<.001**		
Navigation efficacy	(Intercept)	**0.058**	**0.007**	**8.23**	**<.001**	.160	.149
	Instruction recall task	**0.001**	**<0.001**	**3.76**	**<.001**		
Controller motion	(Intercept)	**65541**	**6151**	**10.65**	**<.001**	.006	-.007
	Instruction recall task	-133	191	-0.69	.491		
Total actions	(Intercept)	**593.161**	**52.119**	**11.40**	**<.001**	.099	.087
	Instruction recall task	**-4.633**	**1.621**	**-2.86**	**.006**		
TBPM score	(Intercept)	**3.926**	**0.990**	**3.965**	**<.001**	.030	.017
	Instruction recall task	0.047	0.031	1.523	.132		
Clock checks	(Intercept)	**29.226**	**5.238**	**5.580**	**<.001**	.005	-.008
	Instruction recall task	0.100	0.163	0.617	.539		
EBPM score	(Intercept)	**3.340**	**0.297**	**11.237**	**<.001**	.073	.060
	Instruction recall task	**0.022**	**0.009**	**2.417**	**.018**		

Note. N = 76. The effects that are significant at $p < .05$ are written in bold.

4.4.4 Associations between the EPELI measures and subdomains of parent-reported EF problems

Table 15 shows the correlations between the EPELI and BRIEF measures. The GEC was negatively associated with Task and Navigation efficacies. Of the two indexes that comprise the GEC, the BRI was negatively correlated with Task and Navigation efficacy and positively correlated with Total actions, whereas the MI showed only a negative correlation to Task efficacy. In comparison, the Instruction recall task correlated negatively with the BRI but not with other BRIEF measures (see Supplementary Appendix H in Study III).

Table 15 Correlations between the EPELI and BRIEF measures.

		BRIEF		
		GEC	BRI	MI
EPELI	Total score	-.17	-.15	-.17
	Task efficacy	**-.33 ****	**-.34 ***	**-.29 ***
	Navigation efficacy	**-.29 ***	**-.31 ***	-.24
	Controller motion	.14	.21	.09
	Total actions	.27	**.40 ***	.18
	TBPM score	-.11	-.06	-.12
	Clock checks	.15	.09	.16
	EBPM score	-.22	-.20	-.21

Note. N = 76. The correlations that are significant at $p < .05$ are written in bold. GEC = Global Executive Composite. BRI = Behavioral Regulation Index. MI = Metacognition Index. FDR correction. ** $p < .01$, * $p < .05$.

To study the associations between EPELI and the BRIEF further, linear models with each BRIEF measure as the dependent variable and the EPELI measures as the independent variables were fitted. The best fitting models of these analyses are shown in Table 16. For the GEC and MI, the best models included Task efficacy as the only independent variable, and thus the R^2 of these models is identical to the squares of corresponding correlations in Table 15. For the BRI, the best model was that with Task efficacy and EBPM score as the independent variables.

Table 16 The best fitting models of the analyses with each BRIEF measure as the dependent variable and the EPELI measures as independent variables.

Dependent variable	Independent variable	Estimate (β)	SD	*t*	*p*	R^2	Adj. R^2
BRIEF GEC	**(Intercept)**	**117.061**	**5.645**	**20.78**	**<.001**	.111	.100
	Task efficacy	**-39.115**	**12.834**	**-3.05**	**.003**		
BRIEF BRI	**(Intercept)**	**32.486**	**3.704**	**8.77**	**<.001**	.178	.156
	Task efficacy	**0.015**	**0.004**	**3.53**	**<.001**		
	EBPM score	-1.012	0.724	-1.40	0.166		
BRIEF MI	**(Intercept)**	**76.027**	**4.227**	**17.99**	**<.001**	.086	.073
	Task efficacy	**-25.280**	**9.610**	**-2.63**	**.010**		

Note. N = 76. The effects that are significant at $p < .05$ are written in bold.

4.5 Study IV

4.5.1 EPELI task performance in FSD/HMD and learning effects

The effects of version (FSD/HMD) and time (first/second assessment) on EPELI measures and related descriptive statistics are shown in Table 17. Children achieved higher Total and TBPM scores and Task efficacies in the FSD version with small effect sizes. They made almost twice as many clock checks during the FSD version, which can partly explain the better TBPM performance in that version. To inspect this further, we reran the analysis regarding time monitoring 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 includes TBPM score, we performed additional analysis for Total score without the TBPM tasks. Also here, there were effects of version ($t(67.40) = 2.642$, $p < .01$, $d = 0.32$) and time ($t(67.38) = 6.786$, $p < .001$, $d = 0.83$), which shows that the difference in Total score between the versions is not solely attributable to better TBPM performance in the FSD version.

In the second session, the children acquired higher Total scores (large effect size), TBPM scores (medium effect size), and EBPM scores (small effect size). They also did more actions and navigate more efficiently (small effect sizes for both measures). As there was no difference in Task efficacy between the sessions, it can be reasoned that they did both more relevant and irrelevant actions during the second session compared to the first. In line with this, the number of irrelevant actions (i.e., actions that are not needed for executing the given tasks) increased from the first session to the second ($t(70) = 3.501$, $p < .001$, $d = 0.40$). As learning effects were found in five EPELI measures, we did a post hoc analysis

to enquire if their magnitude was different depending on which version was performed first. There was a larger increase after the HMD version than after the FSD version in Total score (mean change: after HMD 8.42, after FSD 3.18; $t(63.72) = 3.47$, $p < .001$, $d = 0.44$) and TBPM score (mean change: after HMD 3.55, after FSD 1.00; $t(64.50) = 3.395$, $p = .001$, $d = 0.42$). In other measures, version order did not affect the learning effect.

4.5.2 Subjective experiences in FSD/HMD

Table 18 shows the effects of version and time on the Presence questionnaire, which was used to examine differences in participants' subjective experience. The children reported that with the HMD the environment is more involving (small effect size), their experiences felt more consistent with the real world (medium effect size), they can concentrate better on the given tasks (small effect size), and the tasks seemed more interesting (small effect size) than with the FSD version. They also reported more display quality problems for the HMD than FSD version (small effect size), but for both versions, the number of reported problems was small (HMD mean: 2.22; FSD mean: 1.70; on a scale of 1–7). The only difference between the sessions was that the children evaluated the tasks as more interesting after the first session than the second (small effect size). The children reported very few cybersickness symptoms after either version with no difference between the versions (HMD mean: 0.83; FSD mean: 0.56; on a scale of 0–14; $V = 358.5$, $p = .07$).

When asked to compare the versions, most children reported the HMD version as being more realistic (48 out of 51, exact binomial test, $p < .001$) and preferable (36 out of 48, exact binomial test, $p < .001$) than the FSD version. Furthermore, majority of the children reported the HMD version was easier to play, but this difference was not significant (31 out of 49, exact binomial test, $p = .09$).

Table 17 The effect of version (FSD/HMD) and time (1st/2nd session) on the EPELI 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**	**< .001**	**0.44**	**5.814 (0.749)**	**7.765**	**< .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**	**.039**	**0.26**	-0.019 (0.015)	-1.272	.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	.658	-0.05	**0.007 (0.002)**	**2.727**	**.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	.123	-0.20	**43.256 (16.961)**	**2.550**	**.013**	**0.32**
TBPM score (0–13)	6.35 (3.03)	7.67 (2.69)	5.92 (2.91)	8.16 (2.49)	**1.317 (0.371)**	**3.554**	**< .001**	**0.43**	**2.261 (0.371)**	**6.095**	**< .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**	**< .001**	**0.95**	2.918 (3.934)	0.742	.461	0.09
EBPM score (0–6)	4.13 (0.75)	4.29 (0.7)	4.01 (0.77)	4.41 (0.63)	0.157 (0.1)	1.574	.120	0.19	0.396 (0.100)	3.964	<.001	0.48

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

Table 18 The models with the Presence questionnaire items as dependent variables and EPELI version (HMD/FSD) and time (1st/2nd session) as fixed factors.

Question	Descriptive Statistics				Mixed model test statistics					
	Version, mean (SD)		Session, mean (SD)		HMD vs FSD			1st vs 2nd		
	HMD	FSD	1st	2nd	estimate (SD)	*p*	*d*	estimate (SD)	*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)	.298	-0.13	-0.300 (0.200)	.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)**	**.046**	**-0.25**	-0.251 (0.211)	.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)	.434	-0.10	-0.009 (0.215)	.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)**	**<.001**	**-0.50**	-0.074 (0.195)	.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)**	**.002**	**-0.39**	-0.098 (0.168)	.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)	.849	-0.02	-0.045 (0.194)	.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)**	**.014**	**-0.30**	0.078 (0.183)	.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)	.736	-0.03	0.138 (0.107)	.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)	.239	-0.14	0.131 (0.306)	.669	0.05
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)	.050	-0.24	-0.290 (0.149)	.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)**	**.014**	**-0.31**	**-0.526 (0.211)**	**.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)	.451	-0.09	0.052 (0.134)	.697	0.05

Note. N = 72. All questions were answered in a Likert scale with a range of 1–7. The effects that are significant at $p < .05$ are written in bold. Cohen's *d*, effect size. Questions 10–12 are three additional questions that were not in the original Presence Questionnaire 3.0.

4.5.3 Differences between experimenter-supervised laboratory testing and parent-supervised home testing

There were no differences in task performance or perceived presence between the groups who performed FSD-PELI either in laboratory supervised by experiment or at home supervised by parent (Supplementary Tables S3 and S4, Study IV). There were also no differences between these groups in age, handedness, gender, parental education, or family income (Supplementary Table S1, Study IV).

4.5.4 Associations between the EPELI efficacy measures and parent-rated EF deficits

Table 19 contains the correlations between EPELI efficacy measures and the BRIEF across EPELI versions (FSD/HMD) and sessions (first/second). The BRIEF was associated with both Task efficacy and Navigation efficacy on the first session, but not on the second. To shed light on this result, we investigated the association of the BRIEF between the two sessions and found that this correlation was strong (r = .77, $t(67)$ = 9.905, p < .001). To evaluate how carefully the parents had considered their answers, we compared the time used to fill the BRIEF on each session and discovered that parent had used less time on the second session (median time 9.00 mins for the first session and 7.13 mins for the second session, U = 3166, p = .14). When the correlations were calculated separately for both versions but including both sessions, only Navigation efficacy in the HMD version was associated with the BRIEF.

Table 19 Correlations between EPELI efficacy measures and the BRIEF measures across EPELI versions (FSD/HMD) and sessions (first/second).

			BRIEF (GEC)
EPELI measure	session	version	*r*
Task efficacy	first	both	**-.37 ****
Navigation efficacy	first	both	**-.33 ***
Task efficacy	second	both	.03
Navigation efficacy	second	both	-.15
Task efficacy	both	HMD	-.18
Navigation efficacy	both	HMD	**-.37 ****
Task efficacy	both	FSD	-.18
Navigation efficacy	both	FSD	-.11

Note. N = 69–72. The correlations that are significant at p < .05 are written in bold. *** p < .001, ** p < .01, * p < .05, based on p-values with FDR correction.

4.5.5 Inter-version correlations and test-retest stability

Table 20 shows the inter-version and test-retest correlations for the eight EPELI measures and Figure 13 the distributions of the same measures in both versions. Regarding partial correlations between EPELI versions, the highest were in Total score, Task efficacy, and Total actions (.43–.52), followed by Navigation efficacy, Controller motion, TBPM score, EBPM score, and Clock checks (.29–.40). The highest partial correlations between test sessions were those of Total score, Task efficacy, and Total actions (.43–.53), followed by Navigation efficacy, TBPM, EBPM, and Controller motion (.31–.39). Clock checks did not correlate between test sessions. As the effects of version (HMD/FSD) and session (first/second) were analyzed not only for the number of clock checks but also for clock-viewing duration (see section 4.5.1), we also calculated the inter-version and test-retest correlations for this measure and found it to be associated both between test versions (partial $r = .46$, $p < .001$) and test sessions (partial $r = .28$, $p < .05$).

Table 20 EPELI measure intercorrelations between the HMD- and FSD-versions and the first and second sessions.

	HMD vs FSD-version			1st vs 2nd session		
Measure	*r*	ICC[a]	partial *r*[b]	*r*	ICC[a]	partial *r*[d]
Total score	.25	**.23 ***	**.52 ***	**.47 ***	**.32 ***	**.54 ***
Task efficacy	**.47 ***	**.46 ***	**.48 ***	**.45 ***	**.46 ***	**.47 ***
Navigation efficacy	**.36 ****	sing.[c]	**.40 ****	**.39 ****	**.38 ***	**.39 ****
Controller motion	**.33 ****	**.15 ****	**.34 ****	-.16	sing.[c]	**.31 ***
Total actions	**.37 ****	**.38 ****	**.43 ***	**.42 ***	**.41 ***	**.43 ***
TBPM score	.12	.11	**.32 ***	.25	**.19 ***	**.32 ****
Clock checks	**.27 ***	.10	**.29 ***	.15	sing.[c]	.18
EBPM score	.21	.21	**.30 ***	**.29 ***	**.25 ****	**.31 ***

Note. N = 67. The correlations that are significant at $p < 0.05$ are written in bold. *** $p < .001$, ** $p < .01$, * $p < .05$, based on *p*-values with FDR 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. EBPM, event-based prospective memory.

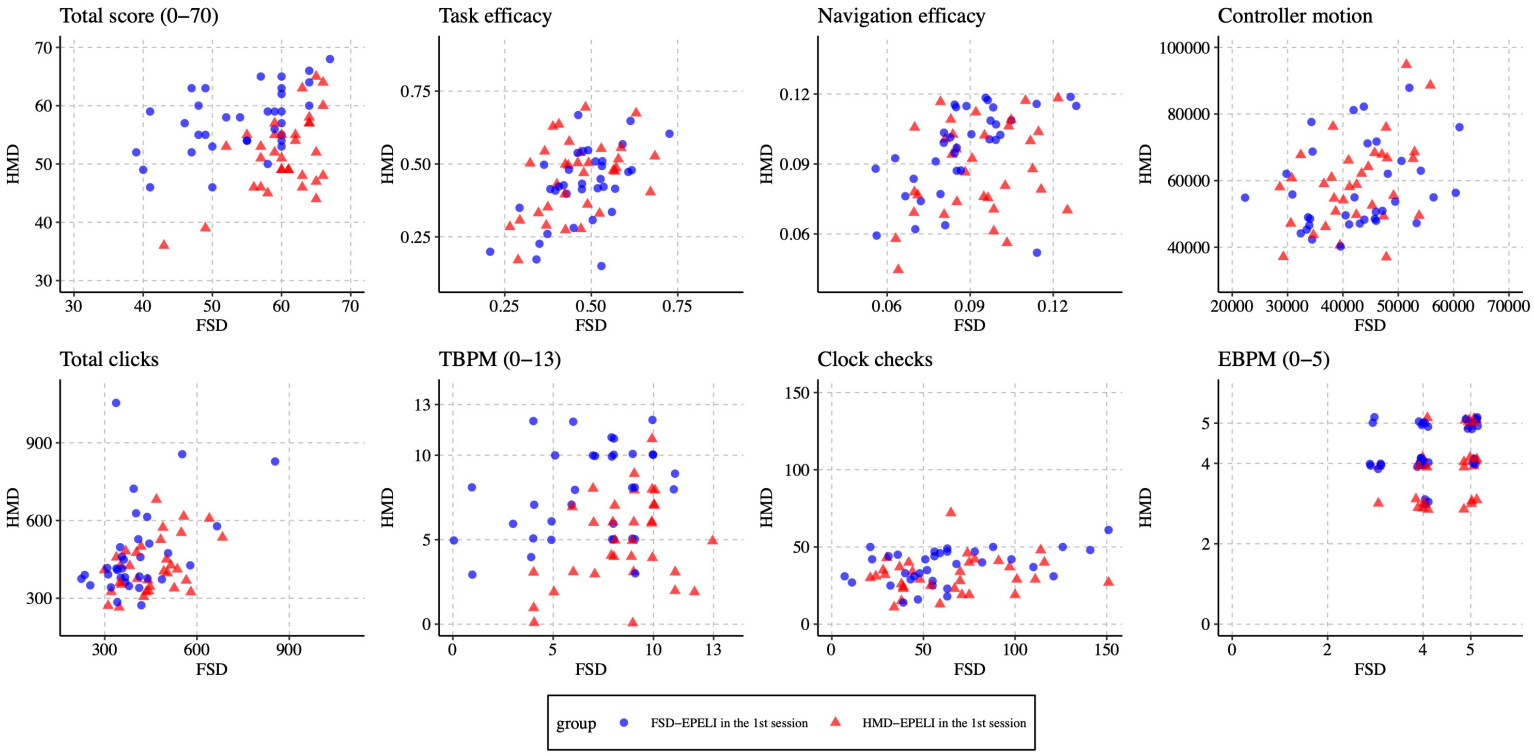


Figure 13 The distributions of EPELI measures per version (HMD/FSD). TBPM, Time-based prospective memory score. EBPM, Event-based prospective memory score. Adapted with permission from Study IV.

4.6 Summary of the main results

The main results of the Dissertation are summarized in Table 21.

Table 21 A summary of the main results.

Area of research	Findings
Concurrent validity via associations between EPELI and traditional performance measures	Study I: observed correlations between EPELI and some conventional attention, PM and EF measures (absolute *r* values [.29–.58]).
Concurrent validity (veridicality) via associations between EPELI and EF ratings	Study I: observed correlations between EPELI and parent-rated EF deficits in a sample consisting of children with ADHD and their controls (absolute *r* values [.36–.57]). Study III: observed correlations between EPELI and parent-rated EF deficits in a sample comprising only typically developing children (absolute *r* values [.29–.40]). Study IV: observed correlations between both the FSD and HMD versions of EPELI and parent-rated EF deficits in a sample comprising typically developing children, but only in the first session (absolute *r* values [.33–.37]).
Predictive, discriminant, and concurrent validity regarding ADHD	Study I: for predictive validity, observed group effects between children with and without ADHD (effect size of η^2_G [.12–28]). Found EPELI to discriminate children with ADHD from typically developing controls (AUC .88). For concurrent validity, observed correlations between EPELI and parent-rated ADHD symptoms (absolute *r* values [.31–.56]) Study II: replicated findings of Study I regarding predictive validity (group effects *Z* = [2.31–3.20]) and discriminant validity (AUC .77).
Eye movement behavior	Study II: found eye movement features to yield higher discriminant validity between children with and without ADHD than the EPELI performance measures introduced in Study I (AUC .92). Successfully used eye movement data to study ADHD-related effects in visual attention.
Usability	Study I: favorable evaluations from the children, no cybersickness symptoms. Study III: favorable evaluations from the children, no cybersickness symptoms, gaming background not found to affect EPELI performance. Study IV: replicated the findings with FSD-EPELI. Found the children to prefer the HMD version and evaluate it to be more immersive.
Reliability in terms of internal consistency	Study III: found adequate to good internal consistency for six out of eight EPELI measures (Cronbach's α [.70–.88]).
Reliability in terms of interversion consistency	Study IV: found all eight EPELI measures to be correlated between the versions (partial *r* values [.29–52]).
Test-retest stability	Study IV: tentative support for test-retest stability with seven out of the eight EPELI measures correlating between the sessions using two different EPELI versions (partial *r* values [.31–54]).

5 Discussion

The overarching aim of this Dissertation was to develop and apply a novel, function-led VR task for the assessment of goal-directed behavior and ADHD symptoms in real-life contexts. Based on its associations with other measures, EPELI successfully quantifies its target phenomena. The results also show that EPELI has good usability and adequate to good psychometric properties, and is considered enjoyable by children. Furthermore, FSD-EPELI, developed for Study IV, allows EPELI to be performed with common laptop and desktop computers, even remotely. Next, the findings are discussed in greater detail.

5.1 Assessing goal-directed behavior with EPELI

5.1.1 Associations between conventional performance measures and EPELI

As conventional construct-driven performance measures originate from various frameworks with premises that are very different from those of the function-led design approach (see Burgess et al., 2006), these tasks may not be the optimal benchmark for new function-led paradigms. Moreover, tasks that simulate everyday functions should logically, like the actual everyday functions, utilize multiple cognitive processes, whereas in many construct-driven tests the rationale is to quantify only one or perhaps a few cognitive domains at a time. Even so, probing the associations between a new test and well-established ones can provide important insights into concurrent validity.

Study I revealed a few associations between some commonly used performance measures and the five EPELI measures included in that study (Total score, Task efficacy, Navigation efficacy, Controller motion, and Total actions). Supporting the concurrent validity of EPELI, links between EPELI and conventional PM measures were observed (Table 9). As EPELI Total score is essentially a PM measure, it is not surprising that it correlated with PM performance in the Cruiser task. The number of errors in the ongoing task of a complex PM task, HEXE, was associated with all five EPELI measures used in Study I. One potential explanation is that a tendency to engage in less goal-oriented and more exploratory behavior

results in a lower number of correctly performed tasks (Total score), lower behavioral efficacy, and a greater amount of physical movement in EPELI, as well as more error-prone performance in HEXE. Regarding measures of other cognitive processes, EPELI efficacy measures correlated negatively with reaction time variability in the CPT, even though this was not the case with the EPELI Total score. This finding could mean that efficient performance in EPELI is related to the ability to maintain a stable level of performance for a prolonged period, as reaction time variability is expected to reflect fluctuations of performance over time, rather than maximal performance (e.g., Sonuga-Barke & Castellanos, 2007). Performance fluctuations were also observed in the EPELI measures (see section 4.2.1), which suggests that open-ended naturalistic tasks could be employed to study attention dynamics as has been done in virtual classroom studies (e.g., Mangalmurti et al., 2020). Commission errors in the CPT have been regarded as being indicative of impulsivity (e.g., Berger et al., 2017) and were associated with the amount of Controller motion and number of Total actions in EPELI (Table 9). This supports the assumption that these two EPELI measures may also be linked to impulsivity. As a general note, the associations found between EPELI and the conventional measures all follow a logical pattern where more error-prone or slower (in terms of reaction time) performance in the conventional tasks is negatively correlated with the EPELI score and efficacy measures but positively with EPELI Controller motion and Total actions.

Based on the role of episodic memory in PM task outcomes (Kliegel et al., 2008b), we hypothesized that a simple instruction recall task would be associated with EPELI performance. This hypothesis was examined in Studies I and III to reveal the extent to which EPELI performance relies on episodic memory processes. Study I found that in a sample comprising both children with and without ADHD, all five EPELI measures were correlated with the Instruction recall task, in which the children verbally repeat similar instructions that they perform during EPELI (see Supplementary Table 3 of Study I). This finding was elaborated in Study III using all eight EPELI measures and a sample of typically developing children (Table 14). Here, Instruction recall task was associated with five of the eight EPELI measures. Even so, the variability explained by the Instruction recall task was only 6–20 percent of the total variability of these EPELI measures, which underlines the role of other cognitive processes, such as attention and EF, besides memory in EPELI performance.

5.1.2 Associations between EF ratings and EPELI

Given that a key motivation for developing EPELI was to create an ecologically valid measure for goal-directed behavior and related concepts such as EF, associations between EF ratings and EPELI are particularly important in

evaluating how EPELI meets its goals. As expected, Study I found EPELI measures to be associated with parent-rated EF problems (the BRIEF questionnaire) in a sample including children with ADHD and typically developing controls (Table 8). These associations follow a logical pattern. EPELI Total score and efficacy measures are negatively associated with parent-rated EF difficulties, while Controller motion and Total actions, which can be regarded to be indicative of hyperactivity and impulsivity, are positively associated with these difficulties. The strongest correlation, which was observed between EPELI Task efficacy and the BRIEF, is stronger than any of the correlations between the EPELI and the conventional neuropsychological tests except for CPT reaction time variability. In earlier studies, the correlations reported between traditional performance-based and rating measures of EF have generally been low (e.g., Toplak et al., 2013; Pino Muñoz & Arán Filippetti, 2021; see also Barkley, 2012). Noteworthy, the correlation between EPELI Task efficacy and the BRIEF is similar to the strongest of the correlations reported by Toplak and collaborators (2013) between the BRIEF and construct-driven EF tests (see also Barkley, 2012; Pino Muñoz & Arán Filippetti, 2021). Regarding ecological validity, this suggests that EPELI matches the earlier performance-based measures in terms of veridicality, while surpassing them in verisimilitude because of the increased representativeness to everyday life. As the differences in sample characteristics between the studies affect these correlations, this comparison is only tentative. Still, this is a promising finding for the first iteration of a new paradigm and suggests that function-led task design could indeed lead to more ecologically valid EF tasks, also in terms of veridicality. In further studies, correlations between EPELI, the BRIEF, and different sets of conventional EF tasks should be tested within the same sample for any differences in magnitude.

In Study III, we found the associations between EPELI efficacy measures and the BRIEF also to exist in a sample of typically developing children (Table 15). However, EPELI Total score, Controller motion and Total actions were not associated with the BRIEF in this sample. This lack of correlation could be affected by the fact that typically developing children have fewer EF problems than children with ADHD (e.g., Faraone et al, 2021), and as expected, the variability in the BRIEF scores was smaller. Another factor could be that the typically developing children display less hyperactivity and impulsivity as indicated by the corresponding EPELI measures (Controller motion and Total actions). Closer inspection on the BRIEF subdomains revealed that problems of behavioral regulation are associated with EPELI Total actions while the BRIEF total score and metacognitive problems are not, which supports the interpretation that Total actions is indicative of impulsivity.

Based on the findings of Study III, the PM-specific EPELI measures (TBPM score, EBPM score and Clock checks) are not associated with the BRIEF in

typically developing children. It can be speculated that the global measures that reflect the overall task performance may be more sensitive to EF problems than the performance in specific cognitive domains. However, further research should be conducted to confirm this.

In Study III, it was also examined which set of EPELI measures provide highest predictive value for each sum measure in the BRIEF (GEC, BRI, and MI). For the GEC and MI, using only Task efficacy as the sole predictor produced the best fitting model, whereas for the BRI, the best fitting model included Task efficacy and EBPM score as predictors (0). This suggests that in EPELI, behavioral problems are best predicted by low Task efficacy, and, in the case of behavioral regulation, also by more difficulties in cue-triggered PM. It also suggests that with the current set of EPELI measures, multivariate methods do not necessarily provide better predictive power for EF problems.

In Study IV, the analysis between the BRIEF and EPELI included only the EPELI efficacy measures and the BRIEF global score GEC, as these had yielded the strongest associations in Studies I and III. Study IV replicated the findings of Studies I and III with the FSD version, which suggests that associations between EPELI and the BRIEF also exist when a non-immersive version of the task is used (Table 19). However, surprisingly, these associations were only observed in the first but not in the second assessment session. This was the case with both EPELI versions. As the correlation of the BRIEF itself between the sessions was strong, this change was most likely due to a change in children's behavior in EPELI from the first session to the second. One potential explanation is that in the second session, children who do not display EF problems in everyday life are also more likely to resort to non-relevant, extraneous behavior more easily, which would make Task efficacy less representative of these problems in the second assessment. This explanation is supported by the fact that on average children performed more actions during the second assessment as compared to the first. An alternative explanation is that it is the novelty of the task that makes Task efficacy representative of EF problems in the first session. This possibility is consistent with the suggestion that the role of EF is particularly prominent in novel situations (Rabbit, 2004). In any case, this result suggests that EPELI may not work optimally for its intended use in subsequent assessments and that further research with EPELI in test-retest settings is needed. It would be interesting to examine whether EPELI successfully discriminates children with and without ADHD not only in the first but also in subsequent assessments. This analysis could not be done in the scope of this Dissertation, as Study IV included only typically developing children.

5.2 Quantifying ADHD symptoms with HMD-EPELI

5.2.1 Associations between HMD-EPELI measures and ADHD

In Study I, the expected differences between children with versus without ADHD were found in all five EPELI measures included in that study, which supports the predictive validity of EPELI (Figure 4 and Table 4; cf. Figure 6). These findings were successfully replicated in Study II using another sample (Figure 7), which supports the robustness of the results. This is especially important considering that while Study I included clinical participants only with predominantly hyperactive/impulsive subtype or combined subtype of ADHD, Study II also included children with predominantly inattentive subtype of ADHD, making it more representative of the full spectrum of ADHD.

In the multiple regression analysis including five EPELI measures, the discriminant validity of EPELI was on par with the CPT (Figure 5), which has been regarded as the current gold standard cognitive test in the ADHD assessment (Albrecht et al., 2015; Ogundele et al., 2011). When considered in the context of ecological validity, this finding could be interpreted as indicating that, regardless of the relative simplicity of the CPT, both tasks show equal veridicality as they discriminate ADHD children from typically developing controls to the same degree. However, as EPELI comprises varied everyday chores in a typical home environment, while the CPT consists only of pressing a single button as a response to a stream of individual stimuli without anything else, the verisimilitude of EPELI can be argued to surpass that of the CPT. A similar argument has been made of another VR task with ecologically valid contexts, the VR-EAL, as Kourtesis and collaborators (2021) found participants to evaluate the verisimilitude of that task to be higher than that of the compared paper-and-pencil tasks.

The inattention symptoms of ADHD, which comprise a varied set of attention and executive problems (see Table 1), were supposed to be reflected especially in the measure of Task efficacy. Indeed, Task efficacy seemed to capture the general ADHD symptoms to the greatest degree, as it showed strongest associations to ADHD symptoms (the ADHD-RS) and EF problems (the BRIEF) as rated by parent and yielded highest accuracy in classifying children with and without ADHD (Table 8 and Table 5). It should be noted that with the sample size used here (n=76), the difference to the correlations between the other EPELI measures and the BRIEF or the classification accuracies of the other EPELI measures were not statistically significant. Still, based on the results of Studies I and II, efficacy measures, that is, measures that take into account both success in task execution and the amount of irrelevant behavior, show the best potential in quantifying the general ADHD symptoms in a life-like task. Task efficacy represents the relative

percentage of relevant actions out of all actions, and it is linked to selective attention, which is typically defined as focusing on a given target while ignoring irrelevant stimuli. Traditional attention measures typically focus on a specific attentional component at a time in a simplified context. In contrast, EPELI Task efficacy is likely to tap various aspects of participant-environment interaction, such as listening to the instructions and keeping them in mind during task performance, planning and executing the task set, and self-monitoring. As such, Task efficacy can be seen to come closer to the diagnostic criteria of inattention in ADHD than the traditional paradigms. This closer match to the diagnostic definitions was expected to result in higher predictive validity and stronger correlations with the subjective questionnaires for Task efficacy than for the conventional neuropsychological tasks, and such results were indeed observed. In the context of ADHD, Task efficacy was considered a global measure of inattention and it showed robust group differences, was effective in predicting the group status of individual participants and had strong correlations with ratings of ADHD symptoms and EF deficits.

Of the three ADHD symptom dimensions, hyperactivity may be the most straightforward to measure, as most of its criteria are related to physical movement of the individual. Previously, activity levels of persons with ADHD have been quantified using various technologies (see meta-analysis by De Crescenzo et al., 2016). The studies that register participant movement during cognitive tasks usually call for a constant inhibition of movement (e.g., Teicher et al., 1996), even though the ADHD hyperactivity symptoms manifest themselves during everyday life situations in which some movement is often normal or even required. In comparison, EPELI hyperactivity measures quantify typical spontaneous behavior during simulated everyday tasks. The results of Study I showed that children with ADHD clearly display excessive Controller motion and Controller motion variability as compared to their control peers during EPELI. Kofler and colleagues (2016) have suggested that hyperactivity is most clearly observed in cognitive tasks with low levels of stimulation. While this may the case, the present results suggest that ADHD-related hyperactivity can also be objectively quantified during open-ended tasks where the participants are moving freely. Although the previous research on the topic has mostly been on head movements (see, e.g., Mangalmurti et al., 2020; Parsons et al., 2019), we chose Controller motion as the primary hyperactivity measure, since it is more straightforwardly linked to executing actions, while head movements also reflect visual search and could hence also relate to successful task performance. Naturally, either head or hand movements (or both) can be employed in future studies.

The third symptom dimension in ADHD is impulsivity, for which the main measure in EPELI was expected to be Total actions. As expected, the children with ADHD performed a higher number of actions, taken to reflect impulsive

interactions with various irrelevant but attractive objects in the environment (e.g., toys, drums, and TV). The ADHD participants also displayed more fluctuation in the number of actions between the scenarios than the typically developing controls (see section 4.2.1). It can be proposed that these impulsive actions toward attractive stimuli in EPELI could be more representative of the daily situations the ADHD children face, than, for example, the lack of ability to inhibit the response to a non-target letter in a continuous sequence of stimuli in the CPT. To draw a comparison to another paradigm, measuring impulsivity in EPELI can be considered similar to delayed reward tasks where the target that triggers impulsive behavior is tempting (e.g., Dalley & Robbins, 2017). As performing impulsive actions in EPELI carries no penalty, the measure is assumed to reflect typical spontaneous behavior.

Analyzing the change in EPELI performance from scenario-to-scenario revealed some further findings related to ADHD (see section 4.2.1). Contrary to our hypothesis, there was relatively little improvement in Total score during EPELI in either group (ADHD/control; see Supplementary Figure 1 in Study I). Even so, both Task and Navigation efficacy declined during EPELI, and more so in the ADHD group than in the control group. At the same time, the amount of Controller motion and the number of Total actions increased during the task. This may reflect an increase in hyperactivity-impulsivity symptoms explained by a decrease in top-down control (e.g., Mangalmurti et al., 2020). Matching our hypothesis, the ADHD group showed more variability in Controller motion and Total actions. Adding distractions and extraneous objects to the scenarios resulted in lower Total scores, lower efficacies, and higher Controller motion in both groups, but the expected stronger distractor effect for the ADHD children was not observed. One explanation for this is that also the non-distracted task scenarios include all kinds of task-irrelevant but tempting objects that may have distracted children with ADHD more than their typically developing peers. This explanation is supported by the fact that ADHD children displayed less efficient performance throughout EPELI. Also, the distractors in EPELI were constantly present during the distracted scenarios, while many of the previous studies have used instantaneous distractors (see Parsons et al., 2019), which may be more ideal for quantifying distractibility in ADHD children. To test this, such distractors have been embedded to newer EPELI versions that are currently used for data collection.

As a task designed using the function-led approach, EPELI consists of everyday scenarios that allow the measurement of different aspects behavior. To quantify ADHD symptoms in VR, others have used the construct-driven approach, where the task design is guided more by the attempt to measure the major symptoms of inattention, hyperactivity, and impulsivity (Fang et al. 2019; Ryu et al., 2020) rather than to maximize the similarity of the tasks as regards to everyday

situations. The EPELI measures reflecting ADHD symptomology should be regarded to quantify a diverse set of cognitive processes and their interactions. To capture specific ADHD symptom dimensions in even greater precision, new methods could certainly be developed. Further machine learning techniques beyond SVMs used in Study II have been suggested as potential tools for improving prediction accuracy (Orrù et al., 2020). As indicated by the eye movement behavior results of Study II, there are also new measures that can be used to attain a higher discriminative and predictive validity. Finally, the measures presented here could be employed to study also other clinical symptoms, for example, those present in autism spectrum disorder. Moreover, the findings should be replicated with the FSD version.

5.2.2 Links between ADHD and eye movement behavior during EPELI

Besides replicating the group differences that support predictive validity and were first reported in Study I with the first five EPELI measures, Study II displayed additional benefits and possibilities of eye tracking in a naturalistic VR task. Eye tracking data considerably improved the discriminative ability of EPELI, as the SVM classifier based on eye movement measures reached 0.92 AUC (Figure 9), which was significantly higher than what was acquired with the SVM based on five EPELI measures in these data (0.77) and with the SVM based on same eye movement metrics in a conventional visual search task Shoot the Target (0.78). Moreover, eye movement data was used to pursue further research questions that are discussed below.

Analyses of the separate eye movement features showed that children with ADHD demonstrate longer fixations and shorter saccades with lower amplitudes than their typically developing peers (Table 10 and Figure 8). Group differences in these parameters have been reported in earlier eye tracking studies with restricted experimental tasks that do not allow active interaction with the environment (Karatekin, 2007; Levantini et al., 2020). Adding to the previous literature, the results of Study II demonstrate that such group differences in eye movement behavior are dependent on the task context. This is an important finding given that in majority of the tasks used in ADHD studies possible behavioral responses are restricted, interaction is limited, and required level of activity is low. Longer fixations in the ADHD group may reflect inefficient and superficial information processing. This explanation is supported by computational modelling work on decision-making in ADHD (Metin et al., 2013; Ziegler et al., 2016) and consistent with the absence of group differences in fixations to relevant objects in Study II (Figure 10). The reasoning here is that shorter fixations to target objects for children with ADHD would presumably have indicated more superficial

processing of target objects. Saccades are used to orient the gaze towards objects that are relevant for ongoing behavior (Eisenberg & Zacks, 2016), and thus their role in inefficient visual search can be argued even more convincingly (Boot et al., 2006). In congruence with the findings of Study II, fewer saccades during task performance have also previously been linked to successful task performance (Tanke et al., 2021). As saccades becoming longer and fixations shorter can be regarded to reflect the development of visual attention (Tanke et al., 2021; Helo et al., 2014), the results here may indicate that maturation of the visual system is delayed in ADHD (Tanke et al., 2021).

In EPELI, the ADHD-related alterations of eye movements were scattered over the whole task rather than being associated either with specific types of events or objects thought to be related to goal-directed (e.g., relevant vs. irrelevant objects) or with stimulus-driven (e.g., object salience) processes. It should be noted that execution of goal-directed behaviors happens throughout EPELI, also when the gaze is not on target objects. Thus, it is likely that the observed eye movement findings are related to goal-directed behavior, although in this type of dynamic task it seems to be the visual search rather than the depth of processing relevant objects that differentiates children with ADHD from their typically developing peers. Even so, the absence of salience effects during active execution of goal-directed behaviors can indicate that such effects, which can be observed in conditions with isolated and static stimuli are presented, are absent in open-ended tasks where top-down processes may suppress them (Risko & Kingstone, 2015). Despite having observed trends that point at possible small effects of bottom-up processes, we did expect more robust evidence for a larger role of stimulus-driven behaviors in ADHD, and the results were interpreted as not supporting the initial hypotheses on this matter.

5.3 Usability, reliability, and scalability of EPELI

In this Dissertation, the usability of EPELI was evaluated by examining the possible effects of gaming experience to EPELI performance and with self-reports regarding possible cybersickness symptoms and the sense of presence, including potential problems with the interface. Reliability was studied by examining EPELI's internal consistency, inter-version correlations, and test-retest stability. By developing an FSD version of EPELI and using that version in home-based remote assessments, scalability of the task beyond its initial form of researcher-supervised immersive HMD-VR was explored. Together, the results in these three areas provided support for EPELI's applicability to further use, but also revealed some limitations of the current versions.

5.3.1 Gaming background, cybersickness and sense of presence

As gaming background can influence performance in computerized cognitive tasks (e.g., Bediou et al., 2018), it was essential to evaluate its potential effects in EPELI. This was done in Study III, and no effects between regular and non-regular gamers were found in any of the eight measures in HMD-EPELI. It is noteworthy that in the same analyses some robust gender and age differences were found with girls outperforming boys and older children the younger ones (Table 13 and Figure 12). Therefore, even though possible small effects of gaming background may not be detected with the present sample size (n=77), it can be reasoned that based on the current knowledge gaming background should not affect EPELI in a degree that would seriously contaminate the results. Similarly, no performance differences were found between gamers and non-gamers in the VR-EAL, another task with real-life contexts and implemented with immersive VR (Kourtesis et al., 2021), even though the gamers went through its scenes more quickly. Zaidi and collaborators (2018) suggest that the more ergonomic and naturalistic interfaces of the HMD systems could diminish the effects of gaming background on naturalistic VR tasks by making the paradigms easy to use for both gamers and non-gamers. It can also be speculated that different types of tasks are probably affected by gaming expertise to a different degree. Both EPELI and the VR-EAL simulate everyday contexts and situations that do not require fast, immediate actions to surprising events, whereas cognitive tasks that place heavy emphasis on reaction time could be more influenced by gaming, especially fast-paced action games (see Bediou et al., 2018). In terms of usability, the most critical finding was that children with no regular gaming background also found it easy to learn the controls in both FSD- and HMD-EPELI, and there were very few problems with the control interfaces (see Table 18).

As brought up in the Introduction (see section 1.6.1), cybersickness can compromise both the comfort of the participant and ecological validity. Very few cybersickness symptoms were reported after the HMD version in Studies I–III (Table 2). Including both HMD-EPELI and FSD-EPELI, Study IV found no difference in cybersickness symptoms between the versions, and the reported symptoms were very negligible for both. These findings are important, as they suggest that regardless of the hardware, the self-paced tasks and naturalistic home environment in EPELI do not induce adverse effects that could make the children uncomfortable and affect the ecological validity (see Weech et al., 2019). Our findings are in concordance with results showing that new HMDs produce only negligible or no cybersickness symptoms (Kourtesis et al., 2019a). Based on adult data, Kourtesis and collaborators (2019b) suggest that to avoid cybersickness symptoms, the maximum duration of an immersive VR session should be between 55 and 70 minutes. The administration time of the full version of EPELI never

exceeds 35 minutes if done non-stop and therefore clearly stays under that recommendation.

In Study IV, different aspects of self-reported sense of presence were evaluated for both EPELI versions (Table 18). For both versions, the answers of the children to the questions regarding naturalness, involvement, consistency with the real world and immersiveness of EPELI on the higher side of the scale (mean 4.39 or more on the scale of 1–7). The naturalness of the moving mechanism was evaluated somewhat less positively (mean 3.93 for the HMD and 3.74 for the FSD version on the scale of 1–7), but actual interference from the control interface for performing the tasks was reported to be low (mean 1.58 for the HMD and 1.55 for the FSD version on the scale of 1–7). These results could be taken as an additional support for the ecological validity in terms of verisimilitude, but the lack of a reference point to which these values could be compared makes this interpretation tentative. Kourtesis and collaborators (2021) compared adult participants' evaluations of verisimilitude of the VR-EAL and paper-and-pencil tasks for everyday life and found significantly higher ratings for the VR task as compared to the paper-and-pencil tasks. It is reasonable to assume that the findings would be similar in children and in other ecologically valid tasks such as EPELI. However, this cannot be confirmed here as the children were not asked to make such comparisons to any conventional paper-and-pencil tasks. The VR-EAL was also evaluated more pleasant than the paper-and-pencil tasks (Kourtesis et al., 2021). Likewise, the answers the children in Study IV gave to questions about feelings of enthusiasm, how interesting the tasks seemed, and how much effort their put in their performance were on the higher end of the scale (mean 4.56–6.25 on the scale of 1–7).

The analysis regarding version differences in the sense of presence revealed some interesting findings favoring the HMD version, which was evaluated as more involving and more consistent with the real world than the FSD version. The children also reported that in the HMD version, they were better able to concentrate on the given tasks and the tasks seemed more interesting to them. In direct comparison between the versions, HMD-EPELI was evaluated to be more realistic and preferable (see section 4.5.2). The results are in line with previous literature that has compared HMD- and FSD-based VR and found that the HMD provokes a stronger sense of presence (see meta-analysis by Caroux, 2023), a greater arousal of positive emotions (e.g., Tan et al., 2015; Pallavicini et al., 2018; Pallavicini et al., 2019; Pallavicini & Pepe, 2019), and greater enjoyment/user satisfaction (Chang et al., 2020; Shelstad et al., 2017) than the FSD.

5.3.2 Reliability

Reflecting the accuracy of the given instrument, reliability is a key criterion for any psychological measure, but it has not been reported for many of the new VR tasks (e.g., Barnett et al., 2021; Chicchi Giglioli et al., 2021; Ouellet et al., 2018). In this Dissertation, the internal consistency results in Study III provide the most reliable evidence of EPELI's reliability. From Study IV, the correlations between the performances of the two EPELI versions (FSD/HMD) separated by a fairly long delay (i.e., over seven months on average) provide preliminary estimates of EPELI's test-retest stability.

Six out of eight EPELI measures were found to have acceptable internal consistency (Cronbach's α = 0.70–0.88) when all 13 scenarios were used (Table 12). The average consistency for all eight EPELI measures was 0.71. If TBPM and EBPM measures, which are included in Total score, are not accounted for separately, the average consistency was 0.79. A similar level of internal consistency (α = 0.79) was recently reported for the VR-EAL (Kourtesis et al., 2021). Burgess and colleagues (2006) claim that function-led tests can be psychometrically as robust as experimentally derived ones, and the findings regarding the two VR tasks discussed here, EPELI and the VR-EAL, support this argument. The only two measures in EPELI that failed to show acceptable internal consistency were those dedicated to PM. For PM measures, a long inter-item interval is often necessary, which limits the acquisition of a sufficient number of datapoints (McDaniel & Einstein, 2007). This may be the reason why PM measures sometimes fail to reach adequate reliability (see, e.g., Mioni et al., 2015). Given that PM plays an important role in daily functioning of the individual (Einstein and McDaniel, 2005; McDaniel & Einstein, 2007) and is important in gaining autonomy and independence from caregivers (Cottini, 2023; Zuber et al., 2019), future research should continue to pursue more reliable PM measures.

The internal consistency of possible shorter EPELI versions was also examined by dropping out one task scenario at a time from the analyses (see Supplementary Appendix B in Study III). It was found that already with approximately half of the scenarios, the internal consistency was acceptable for some of the measures. For example, a version with seven task scenarios instead of thirteen already yielded acceptable internal consistency for the measures of Task efficacy, Controller motion, and Total actions, while being considerably shorter. Interestingly, these measures are the ones considered best for quantifying the ADHD symptom dimensions of inattention, hyperactivity, and impulsivity (see section 5.2). Even though children seem able to perform the full version without being affected by cybersickness and report enjoying doing so, shorter versions can be beneficial for time-constrained situations. As an example, a shorter version with six or seven scenarios has later been developed to be used in clinical neuropsychological evaluations.

Internal consistency could further be improved based on the results of Study III. For example, dropping out the scenario least consistent with the others increases the Cronbach's α of Total score from 0.70 to 0.73 while making the task shorter. Also, replacing items that seem to be functioning worse than the others is a standard procedure in the development of new test versions. Thus, the internal consistency of EPELI may probably be improved by replacing some scenarios or individual test items (i.e., subtasks) with new ones.

Regarding test-retest stability (section 4.5.5), all participants in Study IV performed both FSD- and HMD-EPELI once separated by rather long delay, over seven months on average. This interval is much longer than those used in many studies that focus on test-retest stability (see, e.g., Calamia et al., 2013). This should be considered when interpreting the results, because the test-retest stability has been found to decrease when the test-retest interval becomes longer (Duff, 2012), as the correlation may be affected not only by measurement error but also by true change. Despite the slight differences in the two EPELI versions and the long interval between the sessions, all EPELI measures except the number of clock checks were correlated between the sessions (partial r = 0.31–0.54). These findings can be compared to some earlier studies with shorter time intervals and tasks of various complexity. Backx and others (2020) reported correlations (ρ = 0.39–0.73) between laboratory and home sessions with one week interval with several tasks from the Cambridge Neuropsychological Test Automated Battery. Using a typical dual-task PM task Cruiser, Zuber and colleagues (2021) found correlations between laboratory and home assessment sessions separated by one week (r = 0.56–0.68) and between two laboratory sessions separated by the same interval (r = 0.66–0.78). It has also been suggested that the complex nature of EF tasks, which recruit multiple cognitive processes, makes them more prone to performance variability compared to simpler tasks (Delis et al., 2004). Logically, this should also apply for more complex PM tasks. In line with this, Mioni and colleagues (2015) found varied correlations (r = 0.13–0.74) for somewhat more complex PM task, the Virtual Week. These reference studies point out that test-retest correlations can vary considerably based on task complexity, type of the measure, and the environment (laboratory/home). Noteworthy, test-retest stability of real-world versions of the MET has not yet been sufficiently studied (Rotenberg et al., 2020), even though the original paradigm was developed over 30 years ago. This and the fact that EPELI measures were not correlated with the BRIEF on the second assessment suggest that more research is needed on using function-led tasks in repeated assessments (see also the discussion in section 5.1.2).

It should be noted that the test-retest correlations in Study IV were acquired with two different versions (HMD/FSD) that are not identical. Furthermore, half of the FSD version assessment were made remotely at home. To make more

reliable test-retest stability estimates that would be comparable to earlier literature, future studies should be pursued by using only one EPELI version at a time in laboratory sessions separated by markedly shorter time interval (e.g., two weeks). The test-retest findings reported here such be regarded as initial but have the benefit of showing that almost all EPELI measures are correlated even across two slightly different versions and such a long delay.

5.3.3 Distinctions between the HMD and FSD versions

The FSD version of EPELI was developed as an alternative to the HMD version that would, given the prevalence of typical computers with FSDs, have the potential to widen the paradigm's scalability, reachability, and cost-effectiveness. The results discussed above regarding usability attest to the applicability of both versions. Furthermore, all eight EPELI measures were correlated between the versions (partial r = 0.29–0.52; Table 20), which suggests that the FSD version can be considered to tap the same phenomena as the original HMD version. There were some small-sized effects that suggest that the FSD is slightly easier, and time monitoring behavior seems to be different in the versions. These differences between the versions raise some interesting questions that are discussed next.

Regarding task performance, the children achieved higher Total and TBPM scores and Task efficacies in the FSD version (Table 17 and Figure 13). Some previous studies have found that even though the HMD may promote a higher sense of presence, the FSD can lead to better performance (Barrett et al., 2022; Makransky et al., 2019). This leads to interesting questions about the effect of immersiveness to task performance. It should be noted that in EPELI, the task instructions are given orally and in a similar way in both versions. The children can look around (but not walk around) while listening to the instructions and may be more tempted to do so with the HMD version, which is evaluated to be more involving. Thus, when using the HMD version, the children may focus less on listening to the instructions and more on irrelevant but appealing visual stimuli. Similarly, during the subsequent execution phase, the more immersive experience of the HMD version could lead them to be more distracted by the irrelevant stimuli. Therefore, a reasonable explanation for these inter-version differences is that the higher immersiveness of the HMD version may make children to be more distracted from the task at hand, which leads to a slightly worse performance. Based on eye tracking findings, Barrett and colleagues (2022) argue that simply looking around could be more fun with the HMD as compared to the FSD, which supports the explanation offered above. This explanation aligns with previous research, which demonstrates that children with ADHD, as well as typically developing children, exhibit longer reaction times (Neguț et al., 2017; Pollak et al., 2010) and omission errors (Pollak et al., 2010) with an HMD version of the CPT

compared to an FSD version of the same task. A potential lack of motivation is unlikely to explain the better performance in the FSD version, as the children reported the tasks as more interesting after the HMD version, and there was no difference in reported effort between the versions. The control interface is different in the two versions, but for several reasons, this is unlikely to cause the performance differences. First, most of the children evaluated the HMD version to be the easier to play, and in general, very few problems were reported regarding the control devices with no differences between the versions. Second, no quick actions or particularly skillful use of the controls are needed to perform well in EPELI, as the task does not place heavy time pressure on the participant. There is enough time to perform all required actions even at a relaxed pace if one keeps focused on the given tasks and avoids getting distracted by the environment. Third, it was ensured that all children had sufficient time to familiarize themselves with the controls during the demo section of each EPELI version and use them with ease afterwards.

The differences between the HMD and FSD versions were particularly clear in time monitoring. The number of clock checks in the FSD version were almost twice the number of those during the HMD version, indicating a large-sized effect. This finding could be attributed to at least three separate phenomena. First, a potential higher visual load with the HMD due to a larger FOV could cause less cognitive resources to be available for time monitoring in that version. This potential explanation is compatible with previous research showing that increasing the difficulty of the ongoing task in a PM dual task can result in less active time monitoring (e.g., Khan et al., 2008). Second, in the FSD version, a white circle where the watch appears is also displayed when the time is not shown. This may serve as an additional time monitoring cue compared to the HMD version. Third, in the FSD version the time can be checked with a single click, whereas in the HMD version, the participant needs to raise or rotate the arm slightly and turn the head towards the hand controller to do so. This could reduce the tendency to actively monitor the time. These three possible accounts do not rule each other out and may all play some role behind the differences in checking the time in the two versions. To disentangle their distinctive influences, further EPELI variations, such as an FSD version that does not have a white circle on the screen as a constant potential reminder about time monitoring, could be implemented in forthcoming studies.

To evaluate the potential of FSD-EPELI for remote use, one should consider the FSD-EPELI results between the groups who performed the task either at the laboratory supervised by a researcher or at home supported by a parent. Between these two groups, no differences were found either in EPELI task performance or subjective presence ratings (see Supplementary Tables S2 and S3 in Study IV). To summarize, Study IV shows that remote home testing with FSD-EPELI has the

potential to produce similar results than laboratory-based measurements with HMD-EPELI. Therefore, the FSD version can improve scalability, reachability, and cost-effectiveness of EPELI without compromising the quality of the results. At the same time, the HMD version was found to be more immersive and it permits the measurement of additional behavioral features like head and eye movements.

5.4 Theoretical comparisons with other tasks

Earlier function-led paradigms have inspired the development of EPELI, and some other tasks, such as the CPT and various PM paradigms, provide comparison points for it. Hence, it is fruitful not only to examine the statistical associations between them, but also to consider differences in task structure, amount of and role played by the stimuli, and possible behavioral responses in EPELI and the other paradigms.

Compared to function-led paradigms like the MET that are executed in real-life environments or with real-life objects (cf. Shallice & Burgess, 1991, Chevignard et al., 2010), VR implementation gives EPELI the benefit of an environment that can be meticulously controlled and in which behavior can be measured with the utmost accuracy. As a downside, even the more immersive HMD version of EPELI can be seen as being less ecologically valid in terms of verisimilitude than these real-life tasks, where, for example, object manipulation and moving around is accomplished with natural physical movements. However, VR has the benefit that it allows very different situations and places to be simulated in a short time without the need to move between any actual real-locations or rearrange contents in the physical environment. The scenarios simulated in EPELI comprise a set of various everyday events and, even though all happen in the same home environment, most of the objects need to be changed between the scenarios as different items are present in different everyday situations. Changing these contents between scenarios could be very laborious to implement using only physical objects, whereas in VR, it happens automatically based on previously planned programming. Introducing some of the distractions could also be very challenging. For example, it is not very probable that an annoying fly that follows the participant could be reliably implemented without digital technology. For adults, there exist several VR-based tasks that have taken inspiration from the original MET (see section 1.6.3), and these tests naturally share the benefits of VR with EPELI.

Some special consideration should also be given to the differences between the CPT and EPELI. In Study I, we observed similar discriminative validity, that is, the capacity to classify children into the ADHD and control groups, for EPELI and the CPT. Thus, even though the CPT is a simple paradigm consisting of

responding with one button to the stream of stimuli and is therefore far distant from the richness and variety of everyday situations, it seems to match the current first version of EPELI in providing discriminative validity for ADHD and to capture some core behavioral phenomenon in ADHD. Thus, it could be argued that, at least for this diagnosis, the CPT provides similar ecological validity in terms of veridicality, even though it does not match EPELI in verisimilitude. The relationship between the CPT and ADHD has been studied extensively (Albrecht et al., 2015). Several CPT versions are available, both for FSDs (see Gualtieri & Johnson, 2005) and immersive HMD-VR (e.g., Rizzo et al., 2009; Iriarte et al., 2016; Climent et al., 2021), making the paradigm more accessible to researchers and clinicians than EPELI for the time being. It is argued here that both tasks have their distinctive strengths and weaknesses, which makes them suited for different applications. These relative pros and cons rise from several key differences between the tasks. First, in the CPT, the task is externally paced – that is, the stimuli are presented at a predesigned rate that the participant cannot control. In EPELI, the participant is free to proceed at his own pace, and only the timing of a few events is fixed (i.e., the cues for EBPM tasks). Second, whereas in the CPT the participant needs only to push a single button after each target stimulus, in EPELI he needs to move around the environment that consists of several rooms and manipulate objects to achieve the given goals. Third, only in EPELI is the participant required to keep track of time while performing the other tasks, in order to perform the time-based task at a given moment. Fourth, whereas the target stimuli in different CPT paradigms belong to some fairly restricted category (e.g., letters, small set of pictures that appear on a whiteboard, single words), in EPELI the target stimuli and cues are more varied and include, for example, inanimate objects of various sizes, animated objects with sounds, such as a TV or a running tap, and unanimated objects with sounds, such as a door buzzer, a radio, and a washing machine with an alarm sound. Fifth, EPELI alternates between phases of listening to task instructions and executing them, whereas a typical CPT paradigm is characterized by a short training phase followed by a longer (e.g., 14 minutes) and monotonous execution phase. Because of these differences, EPELI may provide a richer and more heterogenous sample of everyday behavioral responses, which increases its representativeness (verisimilitude) as regards everyday functions (see Burgess et al., 2006). These rich data could also make EPELI particularly well-suited for some machine learning techniques that have been advocated for a data-driven approach (e.g., Vélez, 2021) and developing more efficient techniques in analyzing the results of psychological experiments (e.g., Orrù et al., 2020). On the other hand, the CPT has the benefit of providing a stringent sample of participant responses to target and non-target stimuli, which could possibly result in more reliable sampling of some behavioral metrics, such as reaction time. This could be beneficial, for

example, in screening concussion-related cognitive symptoms (see, e.g., Lecci et al., 2021). Many of the differences between the CPT and EPELI remain to be studied. For example, in Study II, eye movement metrics were successfully used to improve the classification capacity of EPELI from that obtained with the behavioral EPELI measures. The eye movement metric from the compared visual search task, Shoot the Target, failed to match the discrimination accuracy of EPELI, but this does not rule out the possibility that the eye movement metrics derived during the CPT would fail to do so. This option could be studied to determine whether the excellent classification capability obtained with eye tracking during EPELI is more due to characteristics of the task or simply reflects the measurement method (eye tracking instead of behavioral measures such as the number of actions). As there are already several excellent VR-based versions of the CPT (see 1.6.4), such a comparison would be easy to make.

To summarize, even though direct comparisons can be made between EPELI and the CPT variants in some respects, like classification capacity, the distinct characteristics of the tasks make them complementary rather than ruling each other out. It is noteworthy that the studies in this Dissertation represent the first implementation of EPELI, whereas the introduction of the CPT happened almost 70 years ago (Rosvold et al., 1956) and has been followed by many developments and iterations. It is likely that further iterations of EPELI will prove to be better than the first in numerous aspects, and other function-led VR tasks are bound to extend the possibilities of EPELI in many ways.

EPELI can also be compared to the features of typical PM tasks (Burgess et al., 2002) and popular dual-task paradigm tasks like the Cruiser (Kliegel et al., 2013) that was included in Study I. First, like in typical PM tasks, in EPELI the given tasks cannot be fulfilled immediately. However, the delay between any given instruction and the moment it should be executed is less than two minutes, as the instructions last around 30 seconds and the subsequent execution phase a maximum of 90 seconds. Most PM dual-task paradigms include a filler task between the instructions and the first occasion to perform the given PM task, which lengthens the delay. As a typical example, the filler task of the Cruiser task in Zuber and colleagues (2019) lasted three minutes. Second, the ongoing task prevents continuous rehearsal of the PM task in PM paradigms. This can be likened to what happens in EPELI since, as the participant needs to perform a set of tasks in each scenario, there is no idle time that could be used to memorize the later tasks. Whereas in typical PM paradigms the intention cue does not interfere with the ongoing task, in EPELI all tasks are embedded in the same environment and performing one task can provide cues for the remaining tasks (e.g., seeing a TV to be turned off while cleaning the living room). Third, while typical PM paradigms do not provide feedback about the PM performance, in EPELI the participant sees an animation and hears a sound that signals the successful

performance of a given task. This was implemented to make it unambiguous whether a given task had been correctly performed, which may not always be clear in VR environments. Furthermore, during the task development some children spontaneously commented that this animation and sound were motivating. These comparisons show that EPELI resembles dual-task PM tasks in some respects, but is more proximal to complex PM tasks, such as the Six Element Test (Shallice & Burgess, 1991; for PM use see Burgess et al., 2000; Kliegel et al., 2004) and HEXE (Kliegel et al., 2006).

In one important respect, EPELI is quite different from most conventional neuropsychological tasks: most stimuli in EPELI are irrelevant to the tasks at hand and have the potential to distract the participant. In contrast, many conventional neuropsychological tasks, including those employed in the studies of this Thesis (see section 3.3), mostly contain only stimuli that are relevant to the given tasks. In this sense, EPELI converges much more closely to the demands of everyday life situations that often require us to ignore a great deal of irrelevant stimuli.

5.5 Future directions

After the first EPELI versions used here, several new developments have already taken place. For instance, an adult version has been developed (Jylkkä et al., 2023b) and applied to study ADHD (Jylkkä et al., 2023a) and spontaneous memory strategies in adults (Kangas, 2023; Laine et al., 2023). Another EPELI version that can be used during magnetic resonance imaging has been used to study brain activation during this naturalistic task (Kantonistov, 2023; Tauriainen, 2022). The current children's version has also been applied to study irrelevant behavior (Kasteenpohja, 2023), intraindividual fluctuations of attention (Eräste, 2022), and further ADHD-related phenomena (Puhakka, 2021). Importantly, EPELI has also been piloted in clinical use by at least 15 neuropsychologists in Helsinki University Hospital. Naturally, this transition of new methodology to the clinic is essential for realizing the potential of function-led VR assessments in clinical use (see Parsons et al., 2017). The rationale for pursuing function-led task development for the assessment of goal-directed behavior and the findings of this Dissertation suggest directions for further research. Some of these possibilities are discussed next.

Naturalistic tasks like EPELI, which let the participant interact freely with a stimulus-rich environment and call for multiple objectives to be met, produce data with many possibilities for the operationalization of new measures. In this Thesis, the operationalized measures were mostly quite straightforward, such as the total number of correctly performed subtasks or actions during the whole test. In Study I, the variability of the EPELI measures used in that study were also examined,

and in Study II, eye movement behavior was quantified. Further research should continue to explore how to derive different kinds of potentially informative behavioral correlates from the rich EPELI data. For instance, multivariate methods and machine learning techniques could be applied for this purpose (e.g., Orrù et al., 2020; Vélez, 2021). The further machine learning applications include, for example, using convolutional neural networks with visual data (see, e.g., Jha et al., 2023). Machine learning was already used in Study II by using SVMs with EPELI data to enhance the classification accuracy of the task for discriminating ADHD and typically developing children. Some further areas to cover could be to apply similar methods to other clinical groups such as autism spectrum disorder, to study intraindividual variability in greater detail, and to investigate walking trajectories. The effect of spontaneous memory strategies on EPELI performance has already been examined in adults (Laine et al., 2023), and the same should be attempted with children. It should be noted that even though EPELI is considered to quantify goal-directed behavior, none of the EPELI measures is taken to present a "pure" index of a single underlying cognitive construct, such as EF or PM. As already described in the introduction (see section 1.2.1), naturalistic tasks have been acknowledged to recruit both EF and PM, as close relationships between the performance measures of the two constructs have been identified (e.g., Groot et al., 2002; Kliegel et al., 2003; McDaniel et al., 1999). Each EPELI scenario consists of an encoding and an execution phase that differ in eye movement behavior (see Figure 11) and brain activity (Tauriainen, 2022). Further work should be devoted to developing measures that take advantage of the differences between these two phases.

As it is important to validate function-led measures against measures of ability in the real world (Burgess et al., 2006), more research should be conducted to do so with EPELI. As ratings are subjective and include measurement errors, using multiple respondents could be useful to maximize the accuracy of the benchmarking evaluation. In this Thesis, only the parent reports were used, partly because the onset of the COVID-19 pandemic placed heavy demands on the schools and thus the collection of teacher reports was stopped during Study I. In future, asking not only the parent or legal guardian but also the teacher and the child him/herself to fill in ratings for EF and PM could provide the best benchmark of ecological validity for EPELI. Interestingly, the BRIEF manual states that in the normative sample, the correlation between the parent and teacher reports is rather low (r = .34) compared to the test-retest correlations for each report (r = .86 and .91; Gioia et al., 2000). This suggests that the associations between EPELI and the BRIEF could be substantially different depending on the respondent. In addition, future studies should include larger data sets, which would allow more fine-grained analysis, such as examining the associations between EPELI and the individual BRIEF subscales.

Even though a naturalistic function-led task like EPELI may provide opportunities to sample aspects of goal-directed behavior not accessible by simplified construct-driven tasks, it is far from reasonable to rationalize that any single task would prove to be indicative of the whole array of difficulties that individuals encounter while pursuing their goals in everyday life. Instead, several tasks that simulate different environments and functions should be developed (Burgess et al., 2006; Parsons et al., 2017). As both research test batteries and clinical assessments face practical time constraints, the function-led tasks should be compared against each other to see which combination of tasks would be the most informative in each situation. As EPELI has now proven to be a successful VR tool for quantifying goal-directed behavior in children, further modifications should be attempted. One important addition would be to include scenarios with less precise instructions. EF and therefore goal-directed behavior can be difficult to fully assess with conventional performance-based measures in a typical neuropsychological setting because of the structured nature of the tests and testing situation (see Cripe, 1996). While in the current version of EPELI, the child is given a precise list of tasks to be done, giving less specific instructions, such as "Next, you should tidy up your room", would place greater emphasis on the child's ability to plan and initiate his/her actions, which are obviously important aspects of goal-directed behavior. Of course, such even more open-ended tasks face the challenge of how to objectively quantify the success of task execution. Given the importance of the ability to engage in goal-directed behavior for everyday life and the limitations encountered in its measurement, this challenge may be worth accepting.

From the clinical point of view, one of the greatest assets of EPELI and other function-led measures comes from their resemblance to the everyday situations of interest (see Burgess et al., 2006). This means that instead of first performing a neuropsychological task to measure some hypothetical cognitive construct, for example, quantifying working memory by using the respective tasks from the WISC-IV test battery, then scoring and interpreting the results, and finally explaining to the child and the parents what working memory is, how it could be related to the everyday situations of that particular individual, and what kind of compensatory strategies could be tried, the clinician can simply point to concrete performances observed by both the child and clinician in a life-like task and suggest ways in which possible problems could be better overcome. Furthermore, the children may find it easier to describe their own behavior and strategies in executing a goal-directed task when they are doing one, instead of asking them to imagine and explain what could happen in a situation where they are preparing to go to school but fail to do so in time or at all. These qualities of function-led tasks could prove to be particularly beneficial in rehabilitation settings.

5.6 Conclusion

This Dissertation describes the development process and reports key psychometric qualities of a new function-led task, EPELI. To our knowledge, EPELI is the first HMD-based task for school-aged children that quantifies their goal-directed behavior and ADHD symptoms in open-ended everyday scenarios. It provides rich, well-controlled and objective data about these behaviors and symptoms and can be considered an ecologically valid research tool with adequate to excellent psychometric qualities. In clinical settings, it can be used to complement questionnaires and interview that are influenced by subjective bias. Furthermore, EPELI could potentially facilitate communication with the child, parents and teachers about possible difficulties in these areas, as it allows these problems to be replicated in simulated, observable settings. In all, the findings of this Thesis display the possibilities of function-led VR tasks for the study of human behavior (see Parsons, 2017) and suggest that such tasks may also have remarkable potential in the study of many other neuropsychiatric conditions besides ADHD, and in other simulated contexts than the home environment.

References

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4-18 and 1991 profile*. University of Vermont, Department of Psychiatry.

Ackerman, P. L. (1994). Intelligence, attention, and learning: Maximal and typical performance. In D. K. Detterman (Ed.), *Current topics in human intelligence* (Vol. 4, pp. 1–27). Ablex.

Adams, R., Finn, P., Moes, E., Flannery, K., & Rizzo, A. "Skip." (2009). Distractibility in Attention/Deficit/ Hyperactivity Disorder (ADHD): The Virtual Reality Classroom. *Child Neuropsychology*, *15*(2), 120–135. https://doi.org/10.1080/09297040802169077

Agnew-Blais, J. C., Polanczyk, G. V., Danese, A., Wertz, J., Moffitt, T. E., & Arseneault, L. (2018). Young adult mental health and functional outcomes among individuals with remitted, persistent and late-onset ADHD. *The British Journal of Psychiatry*, *213*(3), 526–534. https://doi.org/10.1192/bjp.2018.97

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

Albrecht, B., Uebel-von Sandersleben, H., Wiedmann, K., & Rothenberger, A. (2015). ADHD history of the concept: The case of the Continuous Performance Test. *Current Developmental Disorders Reports*, *2*(1), 10–22. https://doi.org/10.1007/s40474-014-0035-1

Alderman, N., Burgess, P. W., Knight, C., & Henman, C. (2003). Ecological validity of a simplified version of the multiple errands shopping test. *Journal of the International Neuropsychological Society*, *9*(1), 31–44. https://doi.org/10.1017/S1355617703910046

Allain, P., Foloppe, D. A., Besnard, J., Yamaguchi, T., Etcharry-Bouyx, F., Le Gall, D., Nolin, P., & Richard, P. (2014). Detecting everyday action deficits in Alzheimer's disease using a nonimmersive Virtual Reality Kitchen. *Journal of the International Neuropsychological Society*, *20*(5), 468–477. https://doi.org/10.1017/S1355617714000344

American Psychological Association (2015). Goal-directed behavior. In *APA dictionary of psychology*. Retrieved August 14th, 2023, from https://dictionary.apa.org/goal-directed-behavior

Anderson, P. (2002). Assessment and development of Executive Function (EF) during childhood. *Child Neuropsychology*, *8*(2), 71–82. https://doi.org/10.1076/chin.8.2.71.8724

Andrewes, D. G. (2016). *Neuropsychology: From theory to practice (Second edition)*. Routledge, Taylor & Francis Group.

Armitage, S. G. (1946). An analysis of certain psychological tests used for the evaluation of brain injury. *Psychological monographs*, *60*(1), i.

Armstrong, C. M., Reger, G. M., Edwards, J., Rizzo, A. A., Courtney, C. G., & Parsons, T. D. (2013). Validity of the Virtual Reality Stroop Task (VRST) in active duty military. *Journal of Clinical and Experimental Neuropsychology*, *35*(2), 113–123. https://doi.org/10.1080/13803395.2012.740002

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

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., & Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge Neuropsychological Test Automated Battery: A within-subjects counterbalanced study. *Journal of Medical Internet Research*, *22*(8), e16792. https://doi.org/10.2196/16792

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function: executive function. *Mind, Brain, and Education*, *10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Ballhausen, N., Hering, A., Rendell, P. G., & Kliegel, M. (2019). Prospective memory across the lifespan. In Rummel, J., & McDaniel, M. (Eds.). *Prospective memory* (pp. 135–156). UK: Taylor & Francis.

Barkley, R. A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. *Journal of Abnormal Child Psychology*, *19*(2), 149–178. https://doi.org/10.1007/BF00909976

Barkley, R. A. (2012). *Executive functions: What they are, how they work, and why they evolved*. Guilford Press.

Barkley, R. A. (2014). The assessment of executive functioning using the Barkley Deficits in Executive Functioning Scales. In Goldstein, S., & Naglieri, J. A. (Eds.). *Handbook of executive functioning* (pp. 245–263). New York, NY: Springer New York.

Barkley, R. A., & Murphy, K. R. (2010). Impairment in occupational functioning and adult ADHD: The predictive utility of executive function (EF) ratings versus EF tests. *Archives of Clinical Neuropsychology*, *25*(3), 157–173. https://doi.org/10.1093/arclin/acq014

Barkley, R. A., & Murphy, K. R. (2011). The nature of executive function (EF) deficits in daily life activities in adults with ADHD and their relationship to performance on EF tests. *Journal of Psychopathology and Behavioral Assessment*, *33*(2), 137–158. https://doi.org/10.1007/s10862-011-9217-x

Barnett, M. D., Childers, L. G., & Parsons, T. D. (2021). A Virtual Kitchen Protocol to measure everyday memory functioning for meal preparation. *Brain Sciences*, *11*(5), 571. https://doi.org/10.3390/brainsci11050571

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin*, *144*(1), 77–110. https://doi.org/10.1037/bul0000130

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

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

Berger, I., Slobodin, O., & Cassuto, H. (2017). Usefulness and validity of çontinuous performance tests in the diagnosis of attention-deficit hyperactivity disorder children. *Archives of Clinical Neuropsychology*, *32*(1), 81–93. https://doi.org/10.1093/arclin/acw101

Besnard, J., Richard, P., Banville, F., Nolin, P., Aubin, G., Le Gall, D., Richard, I., & Allain, P. (2016). Virtual reality and neuropsychological assessment: The reliability of a virtual kitchen to assess daily-life activities in victims of traumatic brain injury. *Applied Neuropsychology: Adult*, *23*(3), 223–235. https://doi.org/10.1080/23279095.2015.1048514

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function: development of executive functions. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Bitsko, R. H., Claussen, A. H., Lichstein, J., Black, L. I., Jones, S. E., Danielson, M. L., Hoenig, J. M., Jack, S. P. D., Brody, D. J., Gyawali, S., Maenner, M. J., Warner, M., Holland, K. M., Perou, R., Crosby, A. E., Blumberg, S. J., Avenevoli, S., Kaminski, J. W., & Ghandour, R. M. (2022). *Mental Health Surveillance Among Children—United States, 2013–2019. MMWR supplements*, *71*(2), 1.

Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience*, *12*(12), 752–762. https://doi.org/10.1038/nrn3122

Boot, W. R., Kramer, A. F., Becic, E., Wiegmann, D. A., & Kubose, T. (2006). Detecting transient changes in dynamic displays: The more you look, the less you see. *Human Factors, 48*(4), 759–773.

Bordier, C., Puja, F., & Macaluso, E. (2013). Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging. *Neuroimage*, *67*, 213–226.

Born, S., Kerzel, D., & Theeuwes, J. (2011). Evidence for a dissociation between the control of oculomotor capture and disengagement. *Experimental Brain Research*, *208*(4), 621–631. https://doi.org/10.1007/s00221-010-2510-1

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513–531. https://doi.org/10.1037/0003-066X.32.7.513

Brooks, F. P. (1999). What's real about virtual reality? *IEEE Computer Graphics and Applications*, *19*(6), 16–27. https://doi.org/10.1109/38.799723

Brooks, J., Lodge, R., & White, D. (2017). Comparison of a head-mounted display and flat screen display during a micro-UAV target detection task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1514–1518. https://doi.org/10.1177/1541931213601863

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, *50*(3), 255–272. https://doi.org/10.1037/h0060889

Bucci, M. P., Stordeur, C., Septier, M., Acquaviva, E., Peyre, H., & Delorme, R. (2017). Oculomotor abnormalities in children with attention-deficit/hyperactivity disorder are improved by methylphenidate. *Journal of Child and Adolescent Psychopharmacology*, *27*(3), 274–280. https://doi.org/10.1089/cap.2016.0162

Burgess, P. W., Alderman, N., Forbes, C., Costello, A., M-A.Coates, L., Dawson, D. R., Anderson, N. D., Gilbert, S. J., Dumontheil, I., & Channon, S. (2006). The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. *Journal of the International Neuropsychological Society*, *12*(2), 194–209. https://doi.org/10.1017/S1355617706060310

Burgess, P. W., Gonen-Yaacovi, G., & Volle, E. (2011). Functional neuroimaging studies of prospective memory: What have we learnt so far? *Neuropsychologia*, *49*(8), 2246–2257. https://doi.org/10.1016/j.neuropsychologia.2011.02.014

Burgess, P. W., Scott, S. K., & Frith, C. D. (2003). The role of the rostral frontal cortex (area 10) in prospective memory: a lateral versus medial dissociation. *Neuropsychologia*, *41*(8), 906–918.

Burgess, P. W., Veitch, E., De Lacy Costello, A., & Shallice, T. (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia*, *38*(6), 848–863. https://doi.org/10.1016/S0028-3932(99)00134-7

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: meta-analyses of test–retest correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Caldani, S., Razuk, M., Septier, M., Barela, J. A., Delorme, R., Acquaviva, E., & Bucci, M. P. (2019). The effect of dual task on attentional performance in children with ADHD. *Frontiers in Integrative Neuroscience*, *12*, 67. https://doi.org/10.3389/fnint.2018.00067

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

Canty, A. L., Fleming, J., Patterson, F., Green, H. J., Man, D., & Shum, D. H. K. (2014). Evaluation of a virtual reality prospective memory task for use with individuals with severe traumatic brain injury. *Neuropsychological Rehabilitation*, *24*(2), 238–265. https://doi.org/10.1080/09602011.2014.881746

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

Castellanos, F. X., Marvasti, F. F., Ducharme, J. L., Walter, J. M., Israel, M. E., Krain, A., Pavlovsky, C., & Hommer, D. W. (2000). Executive function oculomotor tasks in girls with ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(5), 644–650. https://doi.org/10.1097/00004583-200005000-00019

Castelvecchi, D. (2016). Low-cost headsets boost virtual reality's lab appeal. *Nature*, *533*(7602), 153–154. https://doi.org/10.1038/533153a

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

Chang, C. W., Li, M., Yeh, S. C., Chen, Y., & 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.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13*(4), 181–197. https://doi.org/10.1023/B:NERV.0000009483.91468.fb

Chaytor, N., Schmitter-Edgecombe, M., & Burr, R. (2006). Improving the ecological validity of executive functioning assessment. *Archives of Clinical Neuropsychology*, *21*(3), 217–227. https://doi.org/10.1016/j.acn.2005.12.002

Cheung, C. H. M., McLoughlin, G., Brandeis, D., Banaschewski, T., Asherson, P., & Kuntsi, J. (2017). Neurophysiological correlates of attentional fluctuation in attention-deficit/hyperactivity disorder. *Brain Topography*, *30*(3), 320–332. https://doi.org/10.1007/s10548-017-0554-2

Chevignard, M. P., Catroppa, C., Galvin, J., & Anderson, V. (2010). Development and evaluation of an ecological task to assess executive functioning post childhood TBI: The children's cooking task. *Brain Impairment*, *11*(2), 125–143. https://doi.org/10.1375/brim.11.2.125

Chicchi Giglioli, I. A., Pérez Gálvez, B., Gil Granados, A., & 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, Behavior, and Social Networking*, cyber.2020.0560. https://doi.org/10.1089/cyber.2020.0560

Choi, B. C. K., & Pak, A. W. P. (2005). A Catalog of Biases in Questionnaires. *Preventing chronic disease*, *2*(1), 13.

Cipresso, P., Albani, G., Serino, S., Pedroli, E., Pallavicini, F., Mauro, A., & Riva, G. (2014). Virtual multiple errands test (VMET): A virtual reality-based tool to detect early executive functions deficit in Parkinson's disease. *Frontiers in Behavioral Neuroscience*, *8*. https://doi.org/10.3389/fnbeh.2014.00405

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

Clancy, T. A., Rucklidge, J. J., & Owen, D. (2010). Road-crossing safety in virtual reality: A comparison of adolescents with and without ADHD. *Journal of Clinical Child & Adolescent Psychology*, *35*(2), 203–215. https://doi.org/10.1207/s15374424jccp3502_4

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. https://doi.org/10.1038/nrn755

Cortese, S., Kelly, C., Chabernaud, C., Proal, E., Di Martino, A., Milham, M. P., & Castellanos, F. X. (2012). Toward systems neuroscience of ADHD: A meta-analysis of 55 fMRI Studies. *American Journal of Psychiatry*, *169*(10), 1038–1055. https://doi.org/10.1176/appi.ajp.2012.11101521

Cottini, M. (2023). Improving children's ability to remember intentions: A literature review on strategies to improve prospective memory during childhood. *Psychological Research*. https://doi.org/10.1007/s00426-023-01834-8

Crawford, J., Smith, G., Maylor, E., Della Sala, S., & Logie, R. (2003). The Prospective and Retrospective Memory Questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample. *Memory*, *11*(3), 261–275. https://doi.org/10.1080/09658210244000027

Cripe, L. I. (1996). The ecological validity of executive function testing. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 171–202). Gr Press/St Lucie Press, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dalley, J. W., & Robbins, T. W. (2017). Fractionating impulsivity: Neuropsychiatric implications. *Nature Reviews Neuroscience*, *18*(3), 158–171. https://doi.org/10.1038/nrn.2017.8

Dawson, D., Anderson, N., Burgess, P., Levine, B., Rewilak, D., Cooper, E., Farrow, S., Krpan, K., Peer, M., & Stuss, D. (2005). Poster 47: Naturalistic assessment of executive function: The Multiple Errands Test. *Archives of Physical Medicine and Rehabilitation*, *86*(10), e17.

Dawson, D. R., & Marcotte, T. D. (2017). Special issue on ecological validity and cognitive assessment. *Neuropsychological Rehabilitation*, *27*(5), 599–602. https://doi.org/10.1080/09602011.2017.1313379

De Crescenzo, F., Licchelli, S., Ciabattini, M., Menghini, D., Armando, M., Alfieri, P., Mazzone, L., Pontrelli, G., Livadiotti, S., Foti, F., Quested, D., & Vicari, S. (2016). The use of actigraphy in the monitoring of sleep and activity in ADHD: A meta-analysis. *Sleep Medicine Reviews*, *26*, 9–20. https://doi.org/10.1016/j.smrv.2015.04.002

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system: Examiner's manual*. San Antonio, TX: The Psychological Corporation.

Delis, D. C., Kramer, J. H., Kaplan, E., & Holdnack, J. (2004). Reliability and validity of the Delis-Kaplan Executive Function System: An update. *Journal of the International Neuropsychological Society*, *10*(2), 301–303. https://doi.org/10.1017/S1355617704102191

Denham, S. A., Bassett, H. H., Sirotkin, Y. S., Brown, C., & Morris, C. S. (2015). "No-o-o-o Peeking": Preschoolers' executive control, social competence, and classroom adjustment. *Journal of Research in Childhood Education*, *29*(2), 212–225. https://doi.org/10.1080/02568543.2015.1008659

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

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

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

Doebel, S. (2020). Rethinking executive function and its development. *Perspectives on Psychological Science*, *15*(4), 942–956. https://doi.org/10.1177/1745691620904771

Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of `data.frame`*. 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 of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

DuPaul, G. J. (Ed.). (1998). *ADHD rating scale-IV: Checklists, norms, and clinical interpretation*. Guilford Press.

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, memory, and cognition*, *16*(4), 717.

Eisenberg, M. L., & Zacks, J. M. (2016). Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, *16*(2), 5. https://doi.org/10.1167/16.2.5

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Erez, N., Weiss, P. L., Kizony, R., & Rand, D. (2013). Comparing performance within a virtual supermarket of children with traumatic brain injury to typically developing children: A pilot study. *OTJR: Occupation, Participation and Health*, *33*(4), 218–227. https://doi.org/10.3928/15394492-20130912-04

Eräste, T. (2022). ADHD:hen liittyvä yksilön sisäinen vaihtelu tarkkaavuudessa. Uusi virtuaalitodellisuuteen pohjautuva luonnollisen kaltainen arviointiympäristö. https://urn.fi/URN:NBN:fi-fe2022070551169

Fang, Y., Han, D., & Luo, H. (2019). A virtual reality application for assessment for attention deficit hyperactivity disorder in school-aged children. *Neuropsychiatric Disease and Treatment, Volume 15*, 1517–1523. https://doi.org/10.2147/NDT.S206742

Faraone, S. V., Banaschewski, T., Coghill, D., Zheng, Y., Biederman, J., Bellgrove, M. A., Newcorn, J. H., Gignac, M., Al Saud, N. M., Manor, I., Rohde, L. A., Yang, L., Cortese, S., Almagor, D., Stein, M. A., Albatti, T. H., Aljoudi, H. F., Alqahtani, M. M. J., Asherson, P., … Wang, Y. (2021). The world federation of ADHD international consensus statement: 208 Evidence-based conclusions about the disorder. *Neuroscience & Biobehavioral Reviews*, S014976342100049X. https://doi.org/10.1016/j.neubiorev.2021.01.022

Fernandez-Ruiz, J., Hakvoort Schwerdtfeger, R. M., Alahyane, N., Brien, D. C., Coe, B. C., & Munoz, D. P. (2020). Dorsolateral prefrontal cortex hyperactivity during inhibitory control in children with ADHD in the

antisaccade task. *Brain Imaging and Behavior*, *14*(6), 2450–2463. https://doi.org/10.1007/s11682-019-00196-3

Finnanger, T. G., Andersson, S., Chevignard, M., Johansen, G. O., Brandt, A. E., Hypher, R. E., Risnes, K., Rø, T. B., & Stubberud, J. (2022). Assessment of executive function in everyday life—Psychometric properties of the Norwegian adaptation of the Children's Cooking Task. *Frontiers in Human Neuroscience*, *15*, 761755. https://doi.org/10.3389/fnhum.2021.761755

Fogel, Y., Rosenblum, S., Hirsh, R., Chevignard, M., & Josman, N. (2020). Daily performance of adolescents with executive function deficits: An empirical study using a complex-cooking task. *Occupational Therapy International*, *2020*, 1–11. https://doi.org/10.1155/2020/3051809

Francés, L., Quintero, J., Fernández, A., Ruiz, A., Caules, J., Fillon, G., Hervás, A., & Soler, C. V. (2022). Current state of knowledge on the prevalence of neurodevelopmental disorders in childhood according to the DSM-5: A systematic review in accordance with the PRISMA criteria. *Child and Adolescent Psychiatry and Mental Health*, *16*(1), 27. https://doi.org/10.1186/s13034-022-00462-1

Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsychological assessment. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 91–112). Boca Raton, FL: St Lucie Press.

Fried, M., Tsitsiashvili, E., Bonneh, Y. S., Sterkin, A., Wygnanski-Jaffe, T., Epstein, T., & Polat, U. (2014). ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Research*, *101*, 62–72. https://doi.org/10.1016/j.visres.2014.05.004

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72–89. https://doi.org/10.1038/s41386-021-01132-0

Fu, W. J., Carroll, R. J., & Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, *21*(9), 1979–1986. https://doi.org/10.1093/bioinformatics/bti294

Fuermaier, A. B. M., Fricke, J. A., de Vries, S. M., Tucha, L., & Tucha, O. (2019). Neuropsychological assessment of adults with ADHD: A Delphi consensus study. *Applied Neuropsychology: Adult*, *26*(4), 340–354. https://doi.org/10.1080/23279095.2018.1429441

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

Garden, S. E., Phillips, L. H., & MacPherson, S. E. (2001). Midlife aging, open-ended planning, and laboratory measures of executive function. *Neuropsychology*, *15*(4), 472–482. https://doi.org/10.1037/0894-4105.15.4.472

Geng, J. J., & DiQuattro, N. E. (2010). Attentional capture by a perceptually salient non-target facilitates target processing through inhibition and rapid rejection. *Journal of Vision*, *10*(6), 5–5. https://doi.org/10.1167/10.6.5

Gibson, J. J. (1970). On the Relation between Hallucination and Perception. *Leonardo*, *3*(4), 425. https://doi.org/10.2307/1572259

Gilboa, Y., Jansari, A., Kerrouche, B., Uçak, E., Tiberghien, A., Benkhaled, O., Aligon, D., Mariller, A., Verdier, V., Mintegui, A., Abada, G., Canizares, C., Goldstein, A., & Chevignard, M. (2019). Assessment of executive functions

in children and adolescents with acquired brain injury (ABI) using a novel complex multi-tasking computerised task: The Jansari assessment of Executive Functions for Children (JEF-C ©). *Neuropsychological Rehabilitation*, *29*(9), 1359–1382. https://doi.org/10.1080/09602011.2017.1411819

Gilboa, Y., Rosenblum, S., Fattal-Valevski, A., Toledano-Alhadef, H., Rizzo, A. (Skip), & Josman, N. (2011). Using a Virtual Classroom environment to describe the attention deficits profile of children with Neurofibromatosis type 1. *Research in Developmental Disabilities*, *32*(6), 2608–2613. https://doi.org/10.1016/j.ridd.2011.06.014

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

Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. (2014). Introduction: A History of executive functioning as a theoretical and clinical construct. In S. Goldstein & J. A. Naglieri (Eds.), *Handbook of Executive Functioning* (pp. 3–12). Springer New York. https://doi.org/10.1007/978-1-4614-8106-5_1

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404–411. https://doi.org/10.1037/h0059831

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Grewe, P., Lahr, D., Kohsik, A., Dyck, E., Markowitsch, H. J., Bien, C. G., Botsch, M., & Piefke, M. (2014). Real-life memory and spatial navigation in patients with focal epilepsy: Ecological validity of a virtual reality supermarket task. *Epilepsy & Behavior*, *31*, 57–66. https://doi.org/10.1016/j.yebeh.2013.11.014

Grodzinsky, G. M., & Barkley, R. A. (1999). Predictive power of frontal lobe tests in the diagnosis of attention deficit hyperactivity disorder. *The Clinical Neuropsychologist*, *13*(1), 12–21. https://doi.org/10.1076/clin.13.1.12.1983

Grodzinsky, G. M., & Diamond, R. (1992). Frontal lobe functioning in boys with attention-deficit hyperactivity disorder. *Developmental Neuropsychology*, *8*(4), 427–445. https://doi.org/10.1080/87565649209540536

Groot, Y. C. T., Wilson, B. A., Evans, J., & Watson, P. (2002). Prospective memory functioning in people with and without brain injury. *Journal of the International Neuropsychological Society*, *8*(5), 645–654. https://doi.org/10.1017/S1355617702801321

Gualtieri, C. T., & Johnson, L. G. (2005). ADHD: Is objective diagnosis possible?. *Psychiatry (Edgmont)*, *2*(11), 44.

Haas, M., Zuber, S., Kliegel, M., & Ballhausen, N. (2020). Prospective memory errors in everyday life: Does instruction matter? *Memory*, *28*(2), 196–203. https://doi.org/10.1080/09658211.2019.1707227

Hall, C. L., Valentine, A. Z., Groom, M. J., Walker, G. M., Sayal, K., Daley, D., & Hollis, C. (2016). The clinical utility of the continuous performance test and objective measures of activity for diagnosing and monitoring ADHD in children: A systematic review. *European Child & Adolescent Psychiatry*, *25*(7), 677–699. https://doi.org/10.1007/s00787-015-0798-x

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

Helo, A., Pannasch, S., Sirri, L., & Rämä, P. (2014). The maturation of eye movement behavior: Scene viewing characteristics in children and adults. *Vision Research*, *103*, 83–91. https://doi.org/10.1016/j.visres.2014.08.006

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8). https://doi.org/10.18637/jss.v042.i08.

Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012). Executive functions and self-regulation. *Trends in Cognitive Sciences*, *16*(3), 174–180. https://doi.org/10.1016/j.tics.2012.01.006

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and psychological measurement*, *60*(4), 523-531.

Holm, S., Häikiö, T., Olli, K., & Kaakinen, J. (2021). Eye movements during dynamic scene viewing are affected by visual attention skills and events of the scene: Evidence from first-person shooter gameplay videos. *Journal of Eye Movement Research*, *14*(2). https://doi.org/10.16910/jemr.14.2.3

Iriarte, Y., Díaz-Orueta, U., Cueto, E., Irazustabarrena, P., Banterla, F., & Climent, G. (2016). AULA—Advanced virtual reality tool for the assessment of attention: normative study in spain. *Journal of Attention Disorders*, *20*(6), 542–568. https://doi.org/10.1177/1087054712465335

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Jha, A., Peterson, J. C., & Griffiths, T. L. (2023). Extracting low-dimensional psychological representations from convolutional neural networks. *Cognitive Science*, *47*(1), e13226. https://doi.org/10.1111/cogs.13226

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

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. http://hdl.handle.net/1721.1/68590

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: A review of our current Understanding. *Neuropsychology Review*, *17*(3), 213–233. https://doi.org/10.1007/s11065-007-9040-z

Jylkkä, J., Ritakallio, L., Merzon, L., Kangas, S., Kliegel, M., Zuber, S., Hering, A., Laine, M., & Salmi, J. (2023a). Assessment of goal-directed behavior and prospective memory in adult ADHD with an online 3D videogame simulating everyday tasks. *Scientific Reports*, *13*(1), 9299. https://doi.org/10.1038/s41598-023-36351-6

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

Kadesjö, B., Janols, L.-O., Korkman, M., Mickelsson, K., Strand, G., Trillingsgaard, A., & Gillberg, C. (2004). The FTF (Five to Fifteen): The development of a parent questionnaire for the assessment of ADHD and comorbid conditions. *European Child & Adolescent Psychiatry*, *13*(S3), iii3–iii13. https://doi.org/10.1007/s00787-004-3002-2

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Developmental Review*, *27*(3), 283–348. https://doi.org/10.1016/j.dr.2007.06.006

Kassambara, A. (2020). *Rstatix: Pipe-friendly framework for basic statistical tests*. https://CRAN.R-project.org/package=rstatix.

Kasteenpohja, K. (2023). EPELI-menetelmän aikainen tarpeeton toiminta ja sen yhteys toiminnanohjauksen taitoihin lapsilla : Vertailu VR-laseilla ja tietokoneella. http://urn.fi/URN:NBN:fi:hulib-202303221554

Kangas, S. (2022). Spontaneous internal strategies in a web-based game of prospective memory and executive performance in everyday living: Adults with ADHD and neurotypical adults. http://urn.fi/URN:NBN:fi:hulib-202212144075

Kantonistov, M. (2023). Aivojen toiminnallinen konnektiivisuus ADHD-diagnoosin saaneilla lapsilla lepotilan, videoiden katselun ja virtuaalitodellisuustehtävän aikan. http://urn.fi/URN:NBN:fi:hulib-202303281585

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, *3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

Kerns, K. A. (2000). The CyberCruiser: An investigation of development of prospective memory in children. *Journal of the International Neuropsychological Society*, *6*(1), 62-70.

Kerns, K. A., & Price, K. J. (2001). An investigation of prospective memory in children With ADHD. *Child Neuropsychology*, *7*(3), 162–171. https://doi.org/10.1076/chin.7.3.162.8744

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

Kim, E., Han, J., Choi, H., Prié, Y., Vigier, T., Bulteau, S., & Kwon, G. H. (2021). Examining the academic trends in neuropsychological tests for executive functions using virtual reality: systematic literature review. *JMIR Serious Games*, *9*(4), e30249. https://doi.org/10.2196/30249

Kim, S. (2015). ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665–674. https://doi.org/10.5351/CSAM.2015.22.6.665

Kingstone, A., Smilek, D., Birmingham, E., Cameron, D., & Bischof, W. F. (2005). Cognitive ethology: Giving real life to attention research. In J. Duncan, L. Phillips, & P. McLeod (Eds.), *Measuring the Mind: Speed, control, and age* (pp. 341–358). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198566427.003.0014

Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive Ethology: A new approach for studying human cognition. *British Journal of Psychology*, *99*(3), 317–340. https://doi.org/10.1348/000712607X251243

Klenberg, L., Jämsä, S., Häyrinen, T., Lahti-Nuuttila, P., & Korkman, M. (2010). The Attention and Executive Function Rating Inventory (ATTEX): Psychometric properties and clinical utility in diagnosing ADHD subtypes. *Scandinavian Journal of Psychology*. https://doi.org/10.1111/j.1467-9450.2010.00812.x

Kliegel, M., Eschen, A., & Thöne-Otto, A. I. T. (2004). Planning and realization of complex intentions in traumatic brain injury and normal aging. *Brain and Cognition*, *56*(1), 43–54. https://doi.org/10.1016/j.bandc.2004.05.005

Kliegel, M., Mahy, C. E. V., Voigt, B., Henry, J. D., Rendell, P. G., & Aberle, I. (2013). The development of prospective memory in young schoolchildren: The impact of ongoing task absorption, cue salience, and cue centrality. *Journal of Experimental Child Psychology*, *116*(4), 792–810. https://doi.org/10.1016/j.jecp.2013.07.012

Kliegel, M., Martin, M., & Moor, C. (2003). Prospective memory and ageing: Is task importance relevant? *International Journal of Psychology*, *38*(4), 207–214. https://doi.org/10.1080/00207590344000132

Kliegel, M., McDaniel, M. A., & Einstein, G. O. (2000). Plan formation, retention, and execution in prospective memory: A new approach and age-related effects. *Memory & Cognition*, *28*(6), 1041–1049. https://doi.org/10.3758/BF03209352

Kliegel, M., Jäger, T., Altgassen, M., & Shum, D. (2008a). Clinical neuropsychology of prospective memory. In M. Kliegel, M. A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (pp. 283–308). Taylor & Francis Group/Lawrence Erlbaum Associates.

Kliegel, M., McDaniel, M. A., & Einstein, G. O. (Eds.). (2008b). *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives*. Lawrence Erlbaum Associates.

Kliegel, M., Ropeter, A., & Mackinlay, R. (2006). Complex prospective memory in children with ADHD. *Child Neuropsychology*, *12*(6), 407–419. https://doi.org/10.1080/09297040600696040

Knight, C., Alderman, N., & Burgess, P. W. (2002). Development of a simplified version of the multiple errands test for use in hospital settings. *Neuropsychological Rehabilitation*, *12*(3), 231–255. https://doi.org/10.1080/09602010244000039

Kofler, M. J., Raiker, J. S., Sarver, D. E., Wells, E. L., & Soto, E. F. (2016). Is hyperactivity ubiquitous in ADHD or dependent on environmental demands? Evidence from meta-analysis. *Clinical Psychology Review*, *46*, 12–24. https://doi.org/10.1016/j.cpr.2016.04.004

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review*, *33*(6), 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY - Second Edition (NEPSY - II)*. San Antonio, TX: Psychological Corporation.

Kothgassner, O. D., & Felnhofer, A. (2020). Does virtual reality help to cut the Gordian knot between ecological validity and experimental control? *Annals*

of the International Communication Association, *44*(3), 210–218. https://doi.org/10.1080/23808985.2020.1792790

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2019a). Technological competence is a pre-condition for effective implementation of virtual reality head mounted displays in human neuroscience: A technological review and meta-analysis. *Frontiers in Human Neuroscience*, *13*, 342. https://doi.org/10.3389/fnhum.2019.00342

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2019b). Validation of the Virtual Reality Neuroscience Questionnaire: Maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Frontiers in Human Neuroscience*, *13*, 417. https://doi.org/10.3389/fnhum.2019.00417

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2021). Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An immersive virtual reality neuropsychological battery with enhanced ecological validity. *Journal of the International Neuropsychological Society*, *27*(2), 181–196. https://doi.org/10.1017/S1355617720000764

Kourtesis, P., Korre, D., Collina, S., Doumas, L. A. A., & 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. *Frontiers in Computer Science*, *1*, 12. https://doi.org/10.3389/fcomp.2019.00012

Kourtesis, P., Linnell, J., Amir, R., Argelaguet, F., & MacPherson, S. E. (2023). Cybersickness in virtual reality questionnaire (CSQ-VR): A validation and comparison against SSQ and VRSQ. *Virtual Worlds*, *2*(1), 16–35. https://doi.org/10.3390/virtualworlds2010002

Kourtesis, P., & 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). *Computers in Human Behavior Reports*, *4*, 100151. https://doi.org/10.1016/j.chbr.2021.100151

Krohn, S., Tromp, J., Quinque, E. M., Belger, J., Klotzsche, F., Rekers, S., Chojecki, P., de Mooij, J., Akbal, M., McCall, C., Villringer, A., Gaebler, M., Finke, C., & Thöne-Otto, A. (2020). Multidimensional evaluation of virtual reality paradigms in clinical neuropsychology: Application of the VR-Check framework. *Journal of Medical Internet Research*, *22*(4), e16724. https://doi.org/10.2196/16724

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Kvavilashvili, L., & Ellis, J. (1996). Varieties of intention: Some distinctions and classifications. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), Prospective memory: Theory and applications (pp. 23–51). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Kvavilashvili, L., & Ellis, J. (2004). Ecological validity and twenty years of real-life/laboratory controversy in memory research: A critical (and historical) review. *History and Philosophy of Psychology*.

Kvavilashvili, L., Kyle, F., & Messer, D. J. (2008). The development of prospective memory in children: Methodological issues, empirical findings, and future directions. In Kliegel, M., McDaniel, M. A., & Einstein, G. O. (Eds.).

(2007). *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives*, pp. 115–140. Psychology Press.

Laine, M., Jylkkä, J., Ritakallio, L., Eräste, T., Kangas, S., Hering, A., Zuber, S., Kliegel, M., Fellman, D., & Salmi, J. (2023). EXPRESS: Spontaneous memory strategies in a videogame simulating everyday memory tasks. *Quarterly Journal of Experimental Psychology*, 17470218231183958. https://doi.org/10.1177/17470218231183958

Larsson, L., Schwaller, A., Nyström, M., & Stridh, M. (2016). Head movement compensation and multi-modal event detection in eye-tracking data for unconstrained head movements. *Journal of Neuroscience Methods*, *274*, 13–26. https://doi.org/10.1016/j.jneumeth.2016.09.005

Lecci, L., Freund, C. T., Ayearst, L. E., Sitarenios, G., Pickett, B., Crews, F. S., Dugan, K., Lange, L., Clark, A., Linz, T., Taravath, S., Williams, M., & Keith, J. (2021). Validating a Short Conners CPT 3 as a screener: Predicting self-reported CDC concussion symptoms in children, adolescents, and adults. *Journal of Pediatric Neuropsychology*, *7*(4), 169–181. https://doi.org/10.1007/s40817-021-00107-9

Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, *21*(1), 59–80. https://doi.org/10.1348/026151003321164627

Levantini, V., Muratori, P., Inguaggiato, E., Masi, G., Milone, A., Valente, E., Tonacci, A., & Billeci, L. (2020). EYES Are the window to the mind: Eye-tracking technology as a novel approach to study clinical characteristics of ADHD. *Psychiatry Research*, *290*, 113135. https://doi.org/10.1016/j.psychres.2020.113135

Levin, H. S., Fletcher, J. M., Kufera, J. A., Harward, H., Lilly, M. A., Mendelsohn, D., Bruce, D., & Eisenberg, H. M. (1996). Dimensions of cognition measured by the tower of London and other cognitive tasks in head-injured children and adolescents. *Developmental Neuropsychology*, *12*(1), 17–34. https://doi.org/10.1080/87565649609540638

Levine, B., Dawson, D., Boutet, I., Schwartz, M. L., & Stuss, D. T. (2000). Assessment of strategic self-regulation in traumatic brain injury: Its relationship to injury severity and psychosocial outcome. *Neuropsychology*, *14*(4), 491–500. https://doi.org/10.1037/0894-4105.14.4.491

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

Liversedge, S., Gilchrist, I., & Everling, S. (Eds.). (2011). *The Oxford handbook of eye movements*. OUP Oxford.

Lo, A. H. Y., Humphreys, M., Byrne, G. J., & Pachana, N. A. (2012). Test-retest reliability and practice effects of the Wechsler Memory Scale-III: Test-retest reliability and practice effect in WMS-III. *Journal of Neuropsychology*, *6*(2), 212–231. https://doi.org/10.1111/j.1748-6653.2011.02023.x

Logie, R. H., Trawley, S., & Law, A. (2011). Multitasking: Multiple, domain-specific cognitive functions in a virtual environment. *Memory & Cognition*, *39*(8), 1561–1574. https://doi.org/10.3758/s13421-011-0120-1

Luckerson, V. (2014). Facebook buying oculus virtual-reality company for $2 billion. Retrieved August, 15th, 2023, from https://www.forbes.com/sites/briansolomon/2014/03/25/facebook-buys-oculus-virtual-reality-gaming-startup-for-2-billion/.

Luo, Y., Weibman, D., Halperin, J. M., & Li, X. (2019). A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD). *Frontiers in Human Neuroscience*, *13*, 42. https://doi.org/10.3389/fnhum.2019.00042

Mahy, C. E. V., Moses, L. J., & Kliegel, M. (2014). The development of prospective memory in children: An executive framework. *Developmental Review*, *34*(4), 305–326. https://doi.org/10.1016/j.dr.2014.08.001

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

Mangalmurti, A., Kistler, W. D., Quarrie, B., Sharp, W., Persky, S., & Shaw, P. (2020). Using virtual reality to define the mechanisms linking symptoms with cognitive deficits in attention deficit hyperactivity disorder. *Scientific Reports*, *10*(1), 529. https://doi.org/10.1038/s41598-019-56936-4

Mangiafico, S. S. (2023). *rcompanion: functions to support extension education program evaluation*. Rutgers Cooperative Extension, New Brunswick, New Jersey. version 2.4.30, https://CRAN.R-project.org/package=rcompanion/.

Mariani, M. A., & Barkley, R. A. (1997). Neuropsychological and academic functioning in preschool boys with attention deficit hyperactivity disorder. *Developmental Neuropsychology*, *13*(1), 111–129. https://doi.org/10.1080/87565649709540671

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.

Marsh, R. L., Hicks, J. L., & Cook, G. I. (2008). On beginning to understand the role of context in prospective memory. In M. Kliegel, A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (pp. 77–100). Taylor & Francis Group/ Lawrence Erlbaum Associates.

Matheis, R. J., Schultheis, M. T., Tiersky, L. A., DeLuca, J., Millis, S. R., & Rizzo, A. (2007). Is learning and memory different in a virtual environment? *The Clinical Neuropsychologist*, *21*(1), 146–161. https://doi.org/10.1080/13854040601100668

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates.

McDaniel, M. A., & Einstein, G. O. (2000). Strategic and automatic processes in prospective memory retrieval: A multiprocess framework. *Applied Cognitive Psychology*, *14*(7). https://doi.org/10.1002/acp.775

McDaniel, M. A., & Einstein, G. O. (2007). *Prospective memory an overview and synthesis of an emerging field*. SAGE.

McDaniel, M. A., Glisky, E. L., Guynn, M. J., Rubin, S. R., & Routhieaux, B. C. (1999). Prospective memory: A neuropsychological study. *Neuropsychology, 13*(1), 103–110. https://doi.org/10.1037/0894-4105.13.1.103

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mehta, T., Mannem, N., Yarasi, N. K., & Bollu, P. C. (2020). Biomarkers for ADHD: The present and future directions. *Current Developmental Disorders Reports*, *7*(3), 85–92. https://doi.org/10.1007/s40474-020-00196-9

Metin, B., Roeyers, H., Wiersema, J. R., van der Meere, J. J., Thompson, M., & Sonuga-Barke, E. (2013). ADHD performance reflects inefficient but not impulsive information processing: A diffusion model analysis. *Neuropsychology*, *27*(2), 193–200. https://doi.org/10.1037/a0031533

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, *32*(5), 541–554. https://doi.org/10.1093/arclin/acx050

Mioni, G., Rendell, P. G., Stablum, F., Gamberini, L., & Bisiacchi, P. S. (2015). Test–retest consistency of Virtual Week: A task to investigate prospective memory. *Neuropsychological Rehabilitation*, *25*(3), 419–447. https://doi.org/10.1080/09602011.2014.941295

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Mohammadhasani, N., Caprì, T., Nucita, A., Iannizzotto, G., & Fabio, R. A. (2020). Atypical visual scan path affects remembering in ADHD. *Journal of the International Neuropsychological Society*, *26*(6), 557–566. https://doi.org/10.1017/S135561771900136X

Morey, R. D., & Rouder, J. N. (2015). *Bayes factor: Computation of Bayes factors for common designs*. https://cran.r-project.org/package=BayesFactor.

Mori, M., MacDorman, K., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Mostofsky, S. H., Lasker, A. G., Cutting, L. E., Denckla, M. B., & Zee, D. S. (2001). Oculomotor abnormalities in attention deficit hyperactivity disorder: a preliminary study. *Neurology*, *57*(3), 423-430.

Mühlberger, A., Jekel, K., Probst, T., Schecklmann, M., Conzelmann, A., Andreatta, M., Rizzo, A. A., Pauli, P., & Romanos, M. (2020). The influence of methylphenidate on hyperactivity and attention deficits in children With ADHD: A Virtual Classroom test. *Journal of Attention Disorders*, *24*(2), 277–289. https://doi.org/10.1177/1087054716647480

Neguț, A., Jurma, A. M., & David, D. (2017). Virtual-reality-based attention assessment of ADHD: ClinicaVR: Classroom-CPT versus a traditional continuous performance test. *Child Neuropsychology*, *23*(6), 692–712. https://doi.org/10.1080/09297049.2016.1186617

Neguț, A., Matu, S.-A., Sava, F. A., & David, D. (2016). Virtual reality measures in neuropsychological assessment: A meta-analytic review. *The Clinical*

Neuropsychologist, *30*(2), 165–184. https://doi.org/10.1080/13854046.2016.1144793

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. Freeman.

Nichols, S. L., & Waschbusch, D. A. (2003). A review of the validity of laboratory cognitive tasks used to assess symptoms of ADHD. *Child Psychiatry and Human Development*, *34*(4), 297–315. https://doi.org/10.1023/B:CHUD.0000020681.06865.97

Nigg, J. T., Sibley, M. H., Thapar, A., & Karalunas, S. L. (2020). Development of ADHD: Etiology, heterogeneity, and early life course. *Annual Review of Developmental Psychology*, *2*(1), 559–583. https://doi.org/10.1146/annurev-devpsych-060320-093413

Nolin, P., Stipanicic, A., Henry, M., Joyal, C. C., & Allain, P. (2012). Virtual reality as a screening tool for sports concussion in adolescents. *Brain Injury*, *26*(13–14), 1564–1573. https://doi.org/10.3109/02699052.2012.698359

Norris, G., & Tate, R. L. (2000). The Behavioural Assessment of the Dysexecutive Syndrome (BADS): Ecological, concurrent and construct validity. *Neuropsychological Rehabilitation, 10*(1), 33–45. https://doi.org/10.1080/096020100389282

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). McGraw-Hill.

Official Statistics of Finland. (2022a). Structure of earnings [e-publica- tion]. ISSN=1799-0092. 2016, Appendix table 1. Average monthly earnings and dispersion of earnings of full-time wage and salary earners in 2016 by age group. Helsinki: Statistics Finland [referred: 3.2.2022]. Access method: http://www.stat.fi/til/pra/2016/pra_2016_2017-09-21_tau_001_en.html

Official Statistics of Finland. (2022b). Educational structure of population [e-publication]. ISSN=2242-2919. Helsinki: Statistics Finland [referred: 3.2.2022b]. Access method: http://www.stat.fi/til/vkour/index_en.html.

Ogundele, M. O., Ayyash, H. F., & Banerjee, S. (2011). Role of computerised continuous performance task tests in ADHD. *Progress in Neurology and Psychiatry*, *15*(3), 8–13. https://doi.org/10.1002/pnp.198

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*, 2970. https://doi.org/10.3389/fpsyg.2019.02970

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

Pallavicini, F., Ferrari, A., Pepe, A., Garcea, G., Zanacchi, A., & Mantovani, F. (2018). Effectiveness of virtual reality survival horror games for the emotional elicitation: Preliminary insights using Resident Evil 7: Biohazard. In M. Antona & C. Stephanidis (Eds.), *Universal Access in Human-Computer Interaction. Virtual, Augmented, and Intelligent Environments* (Vol. 10908, pp. 87–101). Springer International Publishing. https://doi.org/10.1007/978-3-319-92052-8_8

Pallavicini, F., & Pepe, A. (2019). Comparing player experience in video games played in virtual reality or on desktop displays: Immersion, flow, and positive emotions. *Extended Abstracts of the Annual Symposium on*

Computer-Human Interaction in Play Companion Extended Abstracts, 195–210. https://doi.org/10.1145/3341215.3355736

Pallavicini, F., Pepe, A., & 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? *Simulation & Gaming*, *50*(2), 136–159. https://doi.org/10.1177/1046878119831420

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

Palmisano, S., Allison, R. S., Teixeira, J., & Kim, J. (2022). Differences in virtual and physical head orientation predict sickness during active head-mounted display-based virtual reality. *Virtual Reality*. https://doi.org/10.1007/s10055-022-00732-5

Pan, X., & 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. *British Journal of Psychology*, *109*(3), 395–417. https://doi.org/10.1111/bjop.12290

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123. https://doi.org/10.1016/S0042-6989(01)00250-4

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00660

Parsons, T. D., & Barnett, M. (2017). Validity of a newly developed measure of memory: Feasibility study of the Virtual Environment Grocery Store. *Journal of Alzheimer's Disease*, *59*(4), 1227–1235. https://doi.org/10.3233/JAD-170295

Parsons, T. D., Bowerly, T., Buckwalter, J. G., & Rizzo, A. A. (2007). A controlled clinical comparison of attention performance in children with ADHD in a virtual reality classroom compared to standard neuropsychological methods. *Child Neuropsychology*, *13*(4), 363–381. https://doi.org/10.1080/13825580600943473

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

Parsons, T. D., Duffield, T., & Asbee, J. (2019). A comparison of virtual reality classroom continuous performance tests to traditional continuous performance tests in delineating ADHD: A meta-analysis. *Neuropsychology Review*, *29*(3), 338–356. https://doi.org/10.1007/s11065-019-09407-6

Parsons, T. D., & Rizzo, A. A. (2008). Initial validation of a virtual environment for assessment of memory functioning: Virtual reality cognitive performance assessment test. *CyberPsychology & Behavior*, *11*(1), 17–25. https://doi.org/10.1089/cpb.2007.9934

Parsons, T. D., & Rizzo, A. "Skip." (2019). A review of virtual classroom environments for neuropsychological assessment. In A. "Skip" Rizzo & S.

Bouchard (Eds.), *Virtual Reality for Psychological and Neurocognitive Interventions* (pp. 247–265). Springer New York. https://doi.org/10.1007/978-1-4939-9482-3_11

Pedersen, T. L. (2020). *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork

Pennington, B. F., & Ozonoff, S. (1996). Executive functions and developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *37*(1), 51–87. https://doi.org/10.1111/j.1469-7610.1996.tb01380.x

Perone, S., Simmering, V. R., & Buss, A. T. (2021). A dynamical reconceptualization of executive-function development. *Perspectives on Psychological Science*, *16*(6), 1198-1208.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416. https://doi.org/10.1016/j.visres.2005.03.019

Pieri, L., Tosi, G., & Romano, D. (2023). Virtual reality technology in neuropsychological testing: A systematic review. *Journal of Neuropsychology*, jnp.12304. https://doi.org/10.1111/jnp.12304

Pievsky, M. A., & McGrath, R. E. (2018). The neurocognitive profile of attention-deficit/hyperactivity disorder: A review of meta-analyses. *Archives of Clinical Neuropsychology*, *33*(2), 143–157. https://doi.org/10.1093/arclin/acx055

Pino Muñoz, M., & Arán Filippetti, V. (2021). Confirmatory factor analysis of the BRIEF-2 parent and teacher form: Relationship to performance-based measures of executive functions and academic achievement. *Applied Neuropsychology: Child*, *10*(3), 219–233. https://doi.org/10.1080/21622965.2019.1660984

Plancher, G., Tirard, A., Gyselinck, V., Nicolas, S., & Piolino, P. (2012). Using virtual reality to characterize episodic memory profiles in amnestic mild cognitive impairment and Alzheimer's disease: Influence of active and passive encoding. *Neuropsychologia*, *50*(5), 592–602. https://doi.org/10.1016/j.neuropsychologia.2011.12.013

Pollak, Y., Shomaly, H. B., Weiss, P. L., Rizzo, A. A., & Gross-Tsur, V. (2010). Methylphenidate effect in children with ADHD can be measured by an ecologically valid continuous performance test embedded in virtual reality. *CNS Spectrums*, *15*(2), 125–130. https://doi.org/10.1017/S109285290002736X

Porffy, L. A., Mehta, M. A., Patchitt, J., Boussebaa, C., Brett, J., D'Oliveira, T., Mouchlianitis, E., & Shergill, S. S. (2022). A novel virtual reality assessment of functional cognition: Validation study. *Journal of Medical Internet Research*, *24*(1), e27641. https://doi.org/10.2196/27641

Pribram, K. H. (1973). The primate frontal cortex–executive of the brain. In K. H. Pribram & A. R. Luria (Eds.), *Psychophysiology of the frontal lobes* (pp. 293–314). Academic Press.

Puhakka, J. (2021). Using a virtual reality-based tool Epeli for the assessment of ADHD in children. http://urn.fi/URN:NBN:fi:hulib-202106303291

Rabbitt, P. (2004). Introduction: methodologies and models in the study of executive function. In P. Rabbit (Ed.), *Methodology of frontal and executive function* (pp. 9–45). Routledge. https://doi.org/10.4324/9780203344187-5

Rand, D., Kizony, R., Feintuch, U., Katz, N., Josman, N., Rizzo, A. "Skip," & Weiss, P. L. (Tamar). (2005). Comparison of two VR platforms for rehabilitation: video capture versus HMD. *Presence: Teleoperators and Virtual Environments*, *14*(2), 147–160. https://doi.org/10.1162/1054746053967012

Rand, D., Rukan, S. B.-A., (Tamar) Weiss, P. L., & Katz, N. (2009). Validation of the virtual MET as an assessment tool for executive functions. *Neuropsychological Rehabilitation*, *19*(4), 583–602. https://doi.org/10.1080/09602010802469074

Raspelli, S., Pallavicini, F., Carelli, L., Morganti, F., Pedroli, E., Cipresso, P., Poletti, B., Corra, B., Sangalli, D., Silani, V., & Riva, G. (2012). Validating the neuro VR-based virtual version of the Multiple Errands Test: Preliminary Results. *Presence: Teleoperators and Virtual Environments*, *21*(1), 31–42. https://doi.org/10.1162/PRES_a_00077

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

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

Rendell, P. G., & Craik, F. I. M. (2000). Virtual week and actual week: Age-related differences in prospective memory. *Applied Cognitive Psychology*, *14*(7), S43–S62. https://doi.org/10.1002/acp.770

Renison, B., Ponsford, J., Testa, R., Richardson, B., & Brownfield, K. (2012). The ecological and construct validity of a newly developed measure of executive function: The Virtual Library Task. *Journal of the International Neuropsychological Society*, *18*(03), 440–450. https://doi.org/10.1017/S1355617711001883

Revelle, W. (2020). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Riggs, N. R., Jahromi, L. B., Razza, R. P., Dillworth-Bart, J. E., & Mueller, U. (2006). Executive function and the promotion of social–emotional competence. *Journal of Applied Developmental Psychology*, *27*(4), 300–309. https://doi.org/10.1016/j.appdev.2006.04.002

Risko, E. F., & Kingstone, A. (2015). Attention in the wild: Visual attention in complex, dynamic, and social environments. In R. R. Hoffman, P. A. Hancock, M. W. Scerbo, R. Parasuraman, & J. L. Szalma (Eds.), *The Cambridge handbook of applied perception research,* Vol. 1, pp. 466–487). Cambridge University Press. https://doi.org/10.1017/CBO9780511973017.030

Riva, G. (1997). Virtual reality as assessment tool in psychology: Virtual reality in neuro-psycho-physiology, *International Journal of Information Management*, *17*(4), 261-270.

Rizzo, A. A., Bowerly, T., Buckwalter, J. G., Klimchuk, D., Mitura, R., & Parsons, T. D. (2009). A virtual reality scenario for all seasons: *The Virtual Classroom*. *CNS Spectrums*, *11*(1), 35–44. https://doi.org/10.1017/S1092852900024196

Rizzo, A. A., Buckwalter, J. G., & Neumann, U. (1997). Virtual reality and cognitive rehabilitation: a brief review of the future. *The Journal of head trauma rehabilitation*, *12*(6), 1-15.

Rizzo, A. “Skip,” & Koenig, S. T. (2017). Is clinical virtual reality ready for primetime? *Neuropsychology*, *31*(8), 877–899. https://doi.org/10.1037/neu0000405

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 1–8.

Rocke, K., Hays, P., Edwards, D., & Berg, C. (2008). Development of a performance assessment of executive function: The Children’s Kitchen Task assessment. *The American Journal of Occupational Therapy*, *62*(5), 528–537. https://doi.org/10.5014/ajot.62.5.528

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, *20*(5), 343–350. https://doi.org/10.1037/h0043220

Rotenberg, S., Ruthralingam, M., Hnatiw, B., Neufeld, K., Yuzwa, K. E., Arbel, I., & Dawson, D. R. (2020). Measurement properties of the Multiple Errands Test: A systematic review. *Archives of Physical Medicine and Rehabilitation*, *101*(9), 1628–1642. https://doi.org/10.1016/j.apmr.2020.01.019

Roth, R. M., Isquith, P. K., & Gioia, G. A. (2013). Assessment of executive functioning using the Behavior Rating Inventory of Executive Function (BRIEF). In Goldstein, S., & Naglieri, J. A. (Eds.). *Handbook of executive functioning* (pp. 301-331). New York, NY: Springer New York.

Rummel, J., Danner, D., & Kuhlmann, B. G. (2019). The short version of the Metacognitive Prospective Memory Inventory (MPMI-s): Factor structure, reliability, validity, and reference data. *Measurement Instruments for the Social Sciences*, *1*(1), 6. https://doi.org/10.1186/s42409-019-0008-6

Rummel, J., & Kvavilashvili, L. (2022). Current theories of prospective memory and new directions for theory development. *Nature Reviews Psychology*, *2*(1), 40–54. https://doi.org/10.1038/s44159-022-00121-4

Ruse, S. A., Harvey, P. D., Davis, V. G., Atkins, A. S., Fox, K. H., & 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 Research: Cognition*, *1*(1), e21–e26. https://doi.org/10.1016/j.scog.2014.01.004

Ryu, S. H., Oh, S., Lee, S., & Chung, T.-M. (2020). A novel approach to diagnose ADHD using virtual reality. In T. K. Dang, J. Küng, M. Takizawa, & T. M. Chung (Eds.), *Future Data and Security Engineering* (Vol. 12466, pp. 260–272). Springer International Publishing. https://doi.org/10.1007/978-3-030-63924-2_15

Sauzéon, H., Arvind Pala, P., Larrue, F., Wallet, G., Déjos, M., Zheng, X., Guitton, P., & N’Kaoua, B. (2012). The use of virtual reality for episodic memory assessment: Effects of active navigation. *Experimental Psychology*, *59*(2), 99–108. https://doi.org/10.1027/1618-3169/a000131

Sayal, K., Prasad, V., Daley, D., Ford, T., & Coghill, D. (2018). ADHD in children and young people: Prevalence, care pathways, and service provision. *The Lancet Psychiatry*, *5*(2), 175–186. https://doi.org/10.1016/S2215-0366(17)30167-0

Schloss, K. B., Schoenlein, M. A., Tredinnick, R., Smith, S., Miller, N., Racey, C., Castro, C., & Rokers, B. (2021). The UW virtual brain project: An immersive

approach to teaching functional neuroanatomy. *Translational Issues in Psychological Science*, *7*(3), 297–314. https://doi.org/10.1037/tps0000281

Shalev, L., Dody, Y., & Mevorach, C. (2010). Impaired selection- and response-related mechanisms in adult-ADHD. *Journal of Vision*, *10*(7), 284–284. https://doi.org/10.1167/10.7.284

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B*, *298*, 199–209.

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

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of clinical psychiatry*, *59*(20), 22-33.

Shelstad, W. J., Smith, D. C., & Chaparro, B. S. (2017). Gaming on the Rift: How Virtual Reality Affects Game User Satisfaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 2072–2076. https://doi.org/10.1177/1541931213602001

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Shute, G. E., & Huertas, V. (1990). Developmental variability in frontal lobe function. *Developmental Neuropsychology*, *6*(1), 1–11. https://doi.org/10.1080/87565649009540445

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Wetzstein, G. (2018). Saliency in VR: how do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, *24*(4), 1633–1642. https://doi.org/10.1109/TVCG.2018.2793599

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, *109*(3), 431–433. https://doi.org/10.1111/bjop.12305

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

Smith, R. E. (2003). The cost of remembering to remember in event-based prospective memory: Investigating the capacity demands of delayed intention performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 347–361. https://doi.org/10.1037/0278-7393.29.3.347

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00328

Sonuga-Barke, E. J. S., & Castellanos, F. X. (2007). Spontaneous attentional fluctuations in impaired states and pathological conditions: A

neurobiological hypothesis. *Neuroscience & Biobehavioral Reviews*, *31*(7), 977–986. https://doi.org/10.1016/j.neubiorev.2007.02.005

Sonuga-Barke, E. J. S., & Halperin, J. M. (2010). Developmental phenotypes and causal pathways in attention deficit/hyperactivity disorder: Potential targets for early intervention?: Developmental phenotypes and causal pathways in ADHD. *Journal of Child Psychology and Psychiatry*, *51*(4), 368–389. https://doi.org/10.1111/j.1469-7610.2009.02195.x

Spooner, D., & Pachana, N. (2006). Ecological validity in neuropsychological assessment: A case for greater consideration in research with neurologically intact populations. *Archives of Clinical Neuropsychology*, *21*(4), 327–337. https://doi.org/10.1016/j.acn.2006.04.004

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stokes, J. D., Rizzo, A., Geng, J. J., & Schweitzer, J. B. (2022). Measuring attentional distraction in children with ADHD using virtual reality technology with eye-tracking. *Frontiers in Virtual Reality*, *3*, 855895. https://doi.org/10.3389/frvir.2022.855895

Strickland, L., Loft, S., Remington, R. W., & Heathcote, A. (2018). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*, *125*(6), 851–887. https://doi.org/10.1037/rev0000113

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Suh, A., & Prophet, J. (2018). The state of immersive technology research: A literature analysis. *Computers in Human Behavior*, *86*, 77–90. https://doi.org/10.1016/j.chb.2018.04.019

Sutherland, I. E. (1965). The ultimate display. In *Proceedings of the IFIP Congress* (Vol. 2, No. 506-508, pp. 506-508).

Talbot, K.-D. S., Müller, U., & Kerns, K. A. (2018). Prospective memory in children with attention deficit hyperactivity disorder: A review. *The Clinical Neuropsychologist*, *32*(5), 783–815. https://doi.org/10.1080/13854046.2017.1393563

Tamm, L., Narad, M. E., Antonini, T. N., O'Brien, K. M., Hawk, L. W., & Epstein, J. N. (2012). Reaction time variability in ADHD: A review. *Neurotherapeutics*, *9*(3), 500–508. https://doi.org/10.1007/s13311-012-0138-5

Tan, C. T., Leong, T. W., Shen, S., Dubravs, C., & Si, C. (2015). Exploring Gameplay experiences on the Oculus Rift. *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, 253–263. https://doi.org/10.1145/2793107.2793117

Tanke, N., Barsingerhorn, A. D., Boonstra, F. N., & Goossens, J. (2021). Visual fixations rather than saccades dominate the developmental eye movement test. *Scientific Reports*, *11*(1), 1162. https://doi.org/10.1038/s41598-020-80870-5

Tauriainen, S. (2022). Unraveling dynamics of brain activation in ADHD children with a novel open-ended virtual reality task. http://urn.fi/URN:NBN:fi:aalto-202208305244.

Teicher, M. H., Ito, Y., Glod, C. A., & Barber, N. I. (1996). Objective measurement of hyperactivity and attentional problems in ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry*, *35*(3), 334–342. https://doi.org/10.1097/00004583-199603000-00015

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

Theeuwes, J., & Belopolsky, A. V. (2012). Reward grabs the eye: Oculomotor capture by rewarding stimuli. *Vision Research*, *74*, 80–85. https://doi.org/10.1016/j.visres.2012.07.024

Thome, J., Ehlis, A.-C., Fallgatter, A. J., Krauel, K., Lange, K. W., Riederer, P., Romanos, M., Taurines, R., Tucha, O., Uzbekov, M., & Gerlach, M. (2012). Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP task force on biological markers and the World Federation of ADHD. *The World Journal of Biological Psychiatry*, *13*(5), 379–400. https://doi.org/10.3109/15622975.2012.690535

Thorell, L. B., Eninger, L., Brocki, K. C., & Bohlin, G. (2010). Childhood Executive Function Inventory (CHEXI): A promising measure for identifying young children with ADHD? *Journal of Clinical and Experimental Neuropsychology*, *32*(1), 38–43. https://doi.org/10.1080/13803390902806527

Thorell, L. B., & Nyberg, L. (2008). The Childhood Executive Functioning Inventory (CHEXI): A new rating instrument for parents and teachers. *Developmental Neuropsychology*, *33*(4), 536–552. https://doi.org/10.1080/87565640802101516

Thöne-Otto, A. I. T., & Walther, K. (2008). Assessment and treatment of prospective memory disorders in clinical practice. In M. Kliegel, M. A. McDaniel, & G. O. Einstein (Eds.), *Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives* (pp. 321–345). Taylor & Francis Group/Lawrence Erlbaum Associates.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct?: Performance-based and rating measures of EF. *Journal of Child Psychology and Psychiatry*, *54*(2), 131–143. https://doi.org/10.1111/jcpp.12001

Toussaint-Thorin, M., Marchal, F., Benkhaled, O., Pradat-Diehl, P., Boyer, F.-C., & Chevignard, M. (2013). Executive functions of children with developmental dyspraxia: Assessment combining neuropsychological and ecological tests. *Annals of Physical and Rehabilitation Medicine*, *56*(4), 268–287. https://doi.org/10.1016/j.rehab.2013.02.006

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00769

Tseng, P.-H., Cameron, I. G. M., Pari, G., Reynolds, J. N., Munoz, D. P., & Itti, L. (2013). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, *260*(1), 275–284. https://doi.org/10.1007/s00415-012-6631-2

Vélez, J. I. (2021). Machine Learning based Psychology: Advocating for a data-driven approach. *International Journal of Psychological Research*, *14*(1). https://doi.org/10.21500/20112084.5365

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., & Joos, M. (2002). Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 145–156. https://doi.org/10.1016/S1369-8478(02)00013-X

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

Walch, M., Frommel, J., Rogers, K., Schüssel, F., Hock, P., Dobbelstein, D., & Weber, M. (2017). Evaluating VR driving simulation from a player experience perspective. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2982–2989. https://doi.org/10.1145/3027063.3053202

Webb, S. S., Jespersen, A., Chiu, E. G., Payne, F., Basting, R., Duta, M. D., & Demeyere, N. (2021). The Oxford digital multiple errands test (OxMET): Validation of a simplified computer tablet based multiple errands test. *Neuropsychological Rehabilitation*, 1–26. https://doi.org/10.1080/09602011.2020.1862679

Wechsler, David. (2003). *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) administration and scoring manual*. The Psychological Corporation.

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity Disorder: A meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

Wilson, A., Dollman, J., Lushington, K., & Olds, T. (2010). Reliability of the 5-min psychomotor vigilance task in a primary school classroom setting. *Behavior Research Methods*, *42*(3), 754–758. https://doi.org/10.3758/BRM.42.3.754

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

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

World Health Organization. (2016). *International statistical classification of diseases and related health problems* (10th ed.). https://icd.who.int/browse10/2016/en.

Yao, S., & Kim, G. (2019). The effects of immersion in a virtual reality game: Presence and physical activity. In X. Fang (Ed.), *HCI in Games* (Vol. 11595, pp. 234–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-22602-2_18

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

Zelazo, P. D., Carlson, S. M., & Kesek, A. (2008). The development of executive function in childhood. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 553–574). Boston Review.

Ziegler, S., Pedersen, M. L., Mowinckel, A. M., & Biele, G. (2016). Modelling ADHD: A review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. *Neuroscience & Biobehavioral Reviews*, *71*, 633–656. https://doi.org/10.1016/j.neubiorev.2016.09.002

Zuber, S., Haas, M., Framorando, D., Ballhausen, N., Gillioz, E., Künzi, M., & Kliegel, M. (2021). The Geneva Space Cruiser: A fully self-administered online tool to assess prospective memory across the adult lifespan. *Memory*, 1–16. https://doi.org/10.1080/09658211.2021.1995435

Zuber, S., Mahy, C. E. V., & Kliegel, M. (2019). How executive functions are associated with event-based and time-based prospective memory during childhood. *Cognitive Development*, *50*, 66–79. https://doi.org/10.1016/j.cogdev.2019.03.001