

See discussions, stats, and author profiles for this publication at: [https://www.researchgate.net/publication/337758605](https://www.researchgate.net/publication/337758605_Multidimensional_evaluation_of_Virtual_Reality_paradigms_in_clinical_neuropsychology_the_VR-Check_framework?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_2&_esc=publicationCoverPdf)

# [Multidimensional evaluation of Virtual Reality paradigms in clinical neuropsychology: the VR-Check framework](https://www.researchgate.net/publication/337758605_Multidimensional_evaluation_of_Virtual_Reality_paradigms_in_clinical_neuropsychology_the_VR-Check_framework?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_3&_esc=publicationCoverPdf)

**Preprint** · December 2019

DOI: 10.31219/osf.io/zsqbe

CITATIONS

0

READS

155

**13 authors**, including:

[Stephan Krohn](https://www.researchgate.net/profile/Stephan_Krohn?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_5&_esc=publicationCoverPdf)

[Charité Universitätsmedizin Berlin](https://www.researchgate.net/institution/Charite_Universitaetsmedizin_Berlin?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_6&_esc=publicationCoverPdf)

**12** PUBLICATIONS **57** CITATIONS

[Johanne Tromp](https://www.researchgate.net/profile/Johanne_Tromp?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_5&_esc=publicationCoverPdf)

[Max Planck Institute for Human Cognitive and Brain Sciences](https://www.researchgate.net/institution/Max_Planck_Institute_for_Human_Cognitive_and_Brain_Sciences?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_6&_esc=publicationCoverPdf)

**7** PUBLICATIONS **47** CITATIONS

[SEE PROFILE](https://www.researchgate.net/profile/Stephan_Krohn?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_7&_esc=publicationCoverPdf)

[SEE PROFILE](https://www.researchgate.net/profile/Johanne_Tromp?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_7&_esc=publicationCoverPdf)

[Eva M Quinque](https://www.researchgate.net/profile/Eva_Quinque?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_5&_esc=publicationCoverPdf) [University of Leipzig](https://www.researchgate.net/institution/University_of_Leipzig?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_6&_esc=publicationCoverPdf)

**14** PUBLICATIONS **91** CITATIONS

[Julia Belger](https://www.researchgate.net/profile/Julia_Belger?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_5&_esc=publicationCoverPdf) [University of Leipzig](https://www.researchgate.net/institution/University_of_Leipzig?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_6&_esc=publicationCoverPdf)

**9** PUBLICATIONS **3** CITATIONS

[SEE PROFILE](https://www.researchgate.net/profile/Eva_Quinque?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_7&_esc=publicationCoverPdf)

[SEE PROFILE](https://www.researchgate.net/profile/Julia_Belger?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_7&_esc=publicationCoverPdf)

**Some of the authors of this publication are also working on these related projects:**

Digital tools for cognitive assessment and rehabilitation [View project](https://www.researchgate.net/project/Digital-tools-for-cognitive-assessment-and-rehabilitation?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_9&_esc=publicationCoverPdf)



The neural consequences of stress (NECOS) and their association to subjective, endocrine, and autonomic stress markers [View project](https://www.researchgate.net/project/The-neural-consequences-of-stress-NECOS-and-their-association-to-subjective-endocrine-and-autonomic-stress-markers?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_9&_esc=publicationCoverPdf)

All content following this page was uploaded by [Stephan Krohn](https://www.researchgate.net/profile/Stephan_Krohn?enrichId=rgreq-c92deebcd155627521fbd4786386cc7a-XXX&enrichSource=Y292ZXJQYWdlOzMzNzc1ODYwNTtBUzo4MzI2NjM4ODc4MzEwNDFAMTU3NTUzMzk0OTA0MA%3D%3D&el=1_x_10&_esc=publicationCoverPdf) on 05 December 2019.

The user has requested enhancement of the downloaded file.

Multidimensional evaluation of Virtual Reality paradigms in clinical neuropsychology the VR-Check framework

## Stephan Krohn1,21 Johanne Tromp2,31 Eva M Quinque3,41 Julia Belger3,41 Felix Klotzsche2,31 Sophia Rekers1,21 Paul Chojecki51 Mert Akbal2,3,61 Cade McCall71 Arno Villringer2,3,41 Michael Gaebler2,3,\*1 Carsten Finke1,2,\*1 and Angelika Thone-Otto3,4,\*

1Department of Neurology1 Charite-Universitatsmedizin Berlin1 Germany

2Berlin School of Mind and Brain1 Humboldt-Universitat zu Berlin1 Germany 3Max-Planck Institute for Human Cognitive and Brain Sciences1 Leipzig1 Germany 4Clinic for Cognitive Neurology1 University Hospital Leipzig1 Germany 5Fraunhofer Institute for Telecommunications1 Heinrich Hertz Institute1 Berlin1 Germany

6Hochschule der Bildenden Kiinste Saar1 Saarbriicken1 Germany

7Department of Psychology1 University of York1 UK

\*These authors contributed equally to the manuscript.

Author for correspondence: Stephan Krohn

Chariteplatz 1

10117 Berlin, Germany

Tel.: 0049 30 2093 89 774

[stephan.krohn@charite.de](mailto:stephan.krohn@charite.de)

Keywords: Virtual Reality, neuropsychology, cognitive assessment, paradigm design, quality criteria

Abstract

Virtual Reality (VR) represents a key technology of the 21st century, attracting substantial inter- est from a wide range of scientific disciplines. With regard to clinical neuropsychology, a multitude of new VR applications are being developed to overcome the limitations of classical paradigms. In consequence, researchers increasingly face the challenge of systematically evaluating the charac- teristics and quality of VR applications in order to design the optimal paradigm for their specific research question and study population. However, the manifold properties of contemporary VR are not adequately captured by the psychometric quality criteria (i.e., objectivity, reliability, validity) commonly referred to by established test theoretical approaches, highlighting the need for an extended paradigm evaluation framework.

To address this gap, we here propose a multidimensional evaluation framework for VR applica- tions in clinical neuropsychology, summarized as an easy-to-use checklist (VR-Check). This frame- work rests on ten main evaluation dimensions encompassing cognitive domain specificity, ecological relevance, technical feasibility, user feasibility, user motivation, task adaptability, performance quan- tification, immersive capacities, training feasibility, and predictable pitfalls. We show how VR-Check enables systematic and comparative paradigm optimization by illustrating its application in an exem- plary research project on the assessment of spatial cognition and executive functions with immersive VR. This application furthermore demonstrates how the framework allows researchers to identify across-domain tradeoffs, makes deliberate design decisions explicit, and optimizes the allocation of study resources. Complementing recent approaches to standardize clinical VR studies, the VR-Check framework enables systematic and project-specific paradigm optimization for behavioral and cogni- tive research in neuropsychology.

# 1 Introduction

Over the past few decades, Virtual Reality (VR) has emerged as one of the most rapidly advancing technologies of the 21st century, attracting substantial attention from a variety of scientific disciplines including neuroscience. VR may be regarded as an umbrella term subsuming the real-time presentation of a computer-generated environment to a human user. Users perceive the environment through visual or multi-sensory stimulation and interact with it through reciprocal data exchange with the computer system, such that VR represents an advanced form of human-computer interaction [(Aukstakalnis &](#_bookmark7) [Blatner, 1992;](#_bookmark7) [Rizzo & Koenig, 2017).](#_bookmark66) VR can be broadly categorized into non-immersive applications

- 2D screen presentations with interaction devices such as a keyboard or a joystick - and immersive applications that are more complex and require the integration of computers with further devices such as head-mounted displays (HMD), VR controllers, or body-tracking sensors. These immersive systems enable users to experience the virtual environment concealed from the outside world and interact with it based on head or body movements.

In the context of developing paradigms for empirical research, VR provides scientists with a unique combination of extensive design possibilities and strong experimental control. Consequently, VR-based approaches are increasingly pursued in biomedical research and specifically with respect to investigating cognitive function with VR (see Fig. [1).](#_bookmark0) As a result, a fast-growing number of neuropsychological VR paradigms is being developed (e.g., [Lalonde, Henry, Drouin-Germain, Nolin, & Beauchamp, 2013;](#_bookmark43) [Negut,](#_bookmark58) [Matu, Sava, & David, 2016;](#_bookmark58) [Parsons & Rizzo, 2008;](#_bookmark60) [Rizzo et al., 2009,](#_bookmark63) [2000;](#_bookmark64) [Rizzo & Kim, 2005;](#_bookmark65) [Rizzo &](#_bookmark66) [Koenig, 2017;](#_bookmark66) [Rizzo, Schultheis, Kerns, & Mateer, 2004;](#_bookmark67) [Schultheis, Himelstein, & Rizzo, 2002;](#_bookmark70) [Schultheis](#_bookmark71) [& Rizzo,](#_bookmark71) [2001;](#_bookmark71) [Teel, Gay, Johnson, & Slobounov,](#_bookmark80) [2016;](#_bookmark80) [Wiener et al.,](#_bookmark83) [2019),](#_bookmark83) paralleled by decreasing costs of hardware components and increasing availability of open-access software systems for creating new VR paradigms in a customized manner (e.g., [Brookes, Warburton, Alghadier, Mon-Williams, &](#_bookmark13) [Mushtaq, 2019;](#_bookmark13) [Griibel et al., 2016;](#_bookmark26) [Howard, 2018;](#_bookmark28) [Zhao et al](#_bookmark85)., [2018).](#_bookmark85) While these advancements open up many opportunities to investigate the clinical potential of VR, they increasingly present researchers with the challenge of defining the optimal paradigm to answer the research question at hand and leverage the advantages of the technology. Screening the VR literature for suitable paradigms, for instance, how should one evaluate the strengths and weaknesses of a particular paradigm, weigh them against each other, and systematically compare quality across several candidate tasks? Similarly, when developing an experimental VR paradigm de novo, what task features are important to consider in the design process, which qualities should an "ideal" VR task possess, and are there tradeoffs in these qualities on which a deliberate design decision must be made?

Here we propose a pragmatic methodological framework to address these questions. To motivate

our approach, we first review traditional task evaluation based on the traditional psychometric quality criteria. We contrast these endeavors with the extensive degrees of freedom in state-of-the-art VR, illustrating that traditional quality criteria alone are inadequate to capture the manifold properties of VR paradigms comprehensively. To overcome this gap, we propose a general and multidimensional evaluation framework for neuropsychological VR paradigms in form of a checklist ("VR-Check"), and we illustrate the application of this framework in a concrete research project. In the following, we focus on VR paradigms for neuropsychological assessment, rather than rehabilitation or cognitive training paradigms. While many of the VR-Check dimensions will be equally relevant to training and rehabilitation tools, we here avoid a conflation of diagnostic and therapeutic VR applications for clarity.

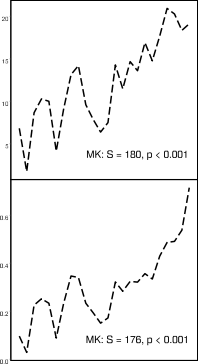
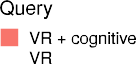




Figure 1: Temporal trends in the biomedical Virtual Reality literature. The PubMed database was searched for unique novel publications in the years 1995-2018 with the queries "Virtual Reality" (VR), "Virtual Reality" AND "cognitive" (VR + cognitive), and "cognitive." Absolute new publication numbers for the former two queries are displayed as bars (search: 09/2019). As absolute publications rose for both the "Virtual Reality" and the "cognitive" query, we computed the respective ratios of publication numbers over time, as shown in the inset. The proportion of annual VR + cognitive PubMed hits over all VR PubMed hits has risen to approximately 20% over the last 20 years, and non-parametric Mann-Kendall trend analysis (MK) indicates a monotonic upward trend of this proportion. A similar temporal trend was observed for the ratio of VR + cognitive over all cognitive PubMed hits, although this proportion remains well under 1%.

## Evaluation criteria in classical neuropsychological tasks

Neuropsychological assessment tools have a longstanding history in clinical neuropsychology, with sev- eral tasks still widely in use more than half a century after their initial presentation (e.g., the Wisconsin Card Sorting Test [(Berg, 1948;](#_bookmark9) [Eling, Derckx, & Maes, 200](#_bookmark20)8), or the Stroop T[est(MacLeod, 1991;](#_bookmark49) [Stroop,](#_bookmark79)

[1935).](#_bookmark79) Early work before the advent of neuroimaging was primarily driven by the aim to measure closely defined cognitive constructs with a clear link to specific brain areas in order to answer diagnostic ques- tions not otherwise solvable at the time [(Chaytor & Schmitter-Edgecombe, 2003).](#_bookmark14) These early tests were predominantly evaluated according to the traditional psychometric quality criteria [(Moosbrugger & Kelava, 2012).](#_bookmark54) In brief, test results had to be independent of the experimenter (objectiv- ity), consistently reproducible over repeated measurements (reliability), and should measure the intended construct (validity) - quality demands that are still widely accepted in cognitive psychology today. With the introduction of neuroimaging into routine diagnostics, however, the mandate for clinical neuropsy- chologists has changed. Rather than helping to identify the neuroetiology, neuropsychologists are now faced with requests to predict and rehabilitate everyday functions, calling for a new type of paradigm tailored to do so [(Chaytor & Schmitter-Edgecombe, 20](#_bookmark14)03). In consequence, the need for an additional evaluation criterion, which better captures the relationship of the neuropsychological paradigm to every- day functioning, has been discussed for some time [(Chaytor & Schmitter-Edgecombe, 2003;](#_bookmark14) [Dawson &](#_bookmark17) [Marcotte, 2017;](#_bookmark17) [Franzen & Wilhelm, 1996;](#_bookmark22) [Parsons, 2015;](#_bookmark59) [Rizzo & Koenig, 2017;](#_bookmark66) [Spooner & Pachana,](#_bookmark78) [2006)).](#_bookmark78) This relationship has been subsumed under the label "ecological relevance" (F[ranzen & Wilhelm](#_bookmark22), [1996;](#_bookmark22) [Rizzo & Koenig, 2017).](#_bookmark66) While in itself still subject to conceptual refinements, ecological relevance is commonly understood to posit that tests should capture the cognitive demands of daily life as closely as possible, resulting in high face validity [(Bornstein, 1996),](#_bookmark11) increased sensitivity to neurorehabilitation, and improved predictive power for everyday functioning [(Dawson & Marcotte,](#_bookmark17) [2017).](#_bookmark17) One landmark publication of a test following these principles is a 1991 paper introducing the Multiple Errands Test (MET, [Shallice & Burgess, 1991)](#_bookmark72) to measure multitasking. The MET comprises a list of shopping-related errands to be performed as a real-life task (i.e., in a real shopping mall). Although the test features high ecological relevance, its limitations include reduced objectivity and reliability due to unforeseeable varia- tions in the real-world mall, high demand for resources to accompany patients in the environment, and - not least - safety issues and inapplicability to patients with more severe disabilities (P[edroli et al., 2016).](#_bookmark61) Although theoretically appealing, real-life tasks have therefore not entered routine neuropsychological assessment and are unlikely to do so due to the lack of control over the test environment. In sum, the search for ecologically relevant, yet experimentally well-controlled tasks is still very much ongoing. In this aspect, VR has the potential to facilitate crucial progress in the field.

## Overcoming limitations of classical tasks with YR-based approaches

Creating a virtual world offers many degrees of freedom: from the environment itself, to the objects in that environment, and even the physics that govern the world. It is therefore possible to design environments

that resemble the real world and its demands much more closely than routine paper-and-pencil tests. At the same time, VR preserves strong control over the experimental conditions (e.g., the existence, type, and frequency of distractors, which are uncontrollable in real-life tasks such as the MET). Similarly, safety concerns of real-world tasks are attenuated by VR paradigms, as patients are not exposed to actual physical dangers (e.g., [Navarro, Llorens, Noe, Ferri, & Alcafiiz, 2013,](#_bookmark56) who used VR to test the act of crossing the street in stroke patients with neglect). Another advantage concerns the increased flexibility of the paradigm development itself: task modifications are implemented computationally, enabling a task design that specifically caters to the study population under question, the research question of interest, or an individual patient's needs. This increased flexibility also illustrates a further limitation of many classical neuropsychological tasks: the lack of parallel versions. In virtual environments, in contrast, parallel task versions are much more easily created by computational modification. Furthermore, routine neuropsychological assessment is highly personnel-dependent, requiring substantial resources in terms of patient assistance and monitoring. Additionally, the evaluation of behavioral performance in classical assessment tasks usually requires time-consuming processing and examination of numeric data (e.g., calculating scores), which then have to be visualized in a graph or table [(Rizzo & Koenig,](#_bookmark66) [2017).](#_bookmark66) In real-life tasks such as the MET, on the other hand, acquisition and evaluation of performance data are even more challenging, as a trained professional has to attend to the patient continuously. VR-based assessment, in contrast, allows for the automatic generation of standardized tests scores and reduces the demand for monitoring resources during assessment. Performance evaluation can be augmented by intuitive feedback to the user (e.g., playback), which may be especially beneficial for certain age or patient groups [(Klinger, Chemin, Lebreton, & Marie,](#_bookmark38) [2006;](#_bookmark38) [Rizzo et al.,](#_bookmark63) [2009,](#_bookmark63) [2004).](#_bookmark67) Finally, the personnel- dependence of traditional approaches constitutes one factor limiting the wide-spread availability of high- quality neuropsychological care, for instance in more rural areas or in patients with restricted mobility. In contrast, VR systems can be employed in patients' homes, offering the long-term perspective of improved ambulatory care and tele-rehabilitation.

## Evaluation criteria in neuropsychological YR paradigms

These advantages of VR raise hopes to ameliorate some of the limitations inherent to classical neuropsy- chological paradigms. However, they also illustrate the multitude of features over which VR paradigms can vary. Currently, neuropsychological VR paradigms are still assessed by using quality criteria which were developed and optimized for methodologically fundamentally different assessments. In general, traditional psychometric quality criteria remain valid for newly developed tests, including VR paradigms. Nonetheless, along with the increased design possibilities of contemporary VR, new evaluation dimensions emerge above and beyond these classical criteria, highlighting the need for

an extended evaluation framework to capture the multidimensional nature of VR applications more ade- quately. Below, we propose such an evaluation framework that allows for a systematic and comparative optimization of VR paradigms in clinical neuropsychology.

# VR-Check: multidimensional evaluation of VR paradigms

The framework rests on ten evaluation dimensions, each comprising several subfeatures. These evaluation criteria are summarized in form of a checklist (VR-Check) in Figure [2A.](#_bookmark3)

## Domain specificity

This evaluation dimension examines how closely the cognitive domain of interest is targeted by the candi- date paradigm. This aspect is especially relevant to VR paradigms, as they differ markedly from classical tasks in both clinical and experimental paradigms: The former usually involves a paper-and-pencil test with task instruction, execution, and evaluation by a trained professional. The latter typically involves the well-controlled presentation of predefined stimuli on a 2D computer screen and the measurement of a predefined set of responses, commonly assessed by interaction devices such as a mouse or a keyboard. In both settings, stimulation is rather unisensory and participants are limited in their ability to act outside the predefined test space. In contrast, VR allows for increased degrees of behavioral freedom, commonly including the liberty to explore the test environment. Compared to classical tasks in neu- ropsychology, VR furthermore permits a much higher level of self-initiated action and interactivity as well as the possibility of multisensory stimulation. While this underscores one particular strength of the technology, these increased degrees of freedom may also recruit other cognitive domains than the one we would like to target. This can make it difficult to ascribe differences in task performance to differences in the cognitive domain under study. Therefore, a VR candidate paradigm should be evaluated on this aspect explicitly. More specifically, it is advisable to a) consider evidence from existing literature that the candidate task will capture the cognitive domain under scrutiny (e.g., are there studies relating the VR task to other assessments whose domain specificity is better established?) and b) to vet the candi- date task for potential domain confounds (e.g., how strongly are visual attention or motor components implicated in solving the task?).

## Ecological relevance

VR enables researchers to simulate real-world scenarios while maintaining a high degree of experimental control. Increasing a task's similarity to the actual challenges encountered by patients in the real world may facilitate diagnostic and rehabilitative approaches that more adequately address the patients' real-life

deficits. This line of thought is commonly subsumed as the potential of VR to increase a task's "ecological validity" [(Dawson & Marcotte,](#_bookmark17) [2017;](#_bookmark17) [Koenig, Crucian, Diinser, Bartneck, & Dalrymple-Alford,](#_bookmark40) [2011;](#_bookmark40) [Parsons, 2015;](#_bookmark59) [Rizzo & Koenig, 2017;](#_bookmark66) [Spooner & Pachana, 2006).](#_bookmark78) As noted above, there is an ongoing debate on what this umbrella term should and should not include on a conceptual level and whether a more fine-grained approach, perhaps along the axes of representativeness and generalizability [(Dawson](#_bookmark17) [& Marcotte](#_bookmark17), [2017),](#_bookmark17) would be beneficial. In opting for the term "ecological relevance," we focus on the patient perspective of everyday functional demands. With "relevant," we deliberately scrutinize potential cognitive deficits of a patient in the domain under study that are likely to translate into a real-world outcomes, such as the ability to function in the real-life environment and perform a real-life action. A candidate task is thus evaluated based on how closely it reflects these demands as encountered by the study population of interest. In consequence, a judgement is made on how relevant the paradigm is to the user's everyday life with respect to a) the virtual environment in which is the task is set, b) the experimental stimuli to which the user is exposed, and c) the activities performed to solve the task (i.e., the user response).

## Technical feasibility

While a candidate paradigm may possess a variety of desirable properties, one may encounter technical limitations when implementing the paradigm in VR. Technical feasibility is especially important to consider if the paradigm is designed de novo or if previously computerized versions of an existing task are not available or incompatible with state-of-the-art VR setups. We therefore evaluate whether the task can be sensibly implemented in VR in general, and whether implementation is compatible with a head-mounted display (HMD), with a 2D display device such as a tablet or a desktop computer, or both. Moreover, it is important to assess whether user interaction and navigation in the virtual world require further input devices such as VR controllers or a mouse, and if so, which input devices are technically feasible. Importantly, technical feasibility of a candidate task may be constrained by project- specific factors such as the necessity of using a particular HMD model, a specific interaction device or an examination room with spatial limitations.

## User feasibility

Candidate paradigms must further be evaluated in terms of feasibility for different user groups. First, is the candidate task expected to be feasible in healthy users, also considering potential differences across different age groups? Second, can one expect it to be viable in the patient population of interest, and are there possibilities to alleviate obstacles in order to maximize patient feasibility? Next, the task is

evaluated on the complexity of the user interaction and navigation in the virtual world: How difficult is it to move and act in the virtual environment? How long will it take healthy and patient users to learn how to carry out the task, and how intuitive are the controls? Furthermore, task duration and attentional demands might limit user feasibility. Therefore, it must be considered how long the task will take on average and whether the target user group can be expected to sufficiently focus on the task. Moreover, user feasibility may be hampered by VR-induced adverse symptoms and effects (VRISE), which are not only important for safety considerations but also because VRISE are likely to confound task performance [(Kourtesis, Collina, Doumas, & MacPherson,](#_bookmark42) [2019).](#_bookmark42) One should therefore judge the paradigm on the likelihood of inducing VRISE such as VR-related kinetosis ("cybersickness"). Not least, it is important to evaluate any ethical concerns the task may implicate, such as the presentation of strong fear-inducing stimuli or safety considerations as mentioned above (see also [Madary and Metzinger (2016](#_bookmark50)) for a detailed review of ethical considerations in VR). While these are relevant aspects to evaluate in any study population, the judgment on what is feasible in the target user group may certainly differ depending on population-specific factors such as health status or age.

## User motivation

Beyond mere feasibility, user motivation is crucial to ensure that participants will engage in the candidate task, especially in repeated application. To optimize user compliance, it is therefore advisable to evaluate the task with regard to factors that may facilitate user motivation. First, it is evaluated to what extent users may be intrinsically motivated to carry out the task due to high expected benefit or face validity of the paradigm. Second, the entertainment factor of the candidate task is judged. Next, one evaluates the possibility of a reward system, both within-session (e.g., a virtual reward for successful task completion) and across-session (e.g., a high score system or advancing to higher levels). Furthermore, we examine the possibility of within- and across-session feedback on the user performance. These features touch upon a gamification approach to cognitive assessment [(Deterding, Dixon, Khaled, & Nacke, 2011;](#_bookmark18) [Lumsden,](#_bookmark48) [Edwards, Lawrence, Coyle, & Munafo, 2016),](#_bookmark48) and this represents one aspect in which VR is particularly capable of playing off its strengths against classical neuropsychological assessments.

## Task adaptability

The ability to adapt the candidate paradigm carries important implications for both clinical and exper- imental settings. First, it is evaluated how easily parallel versions of the candidate task can be created, which represents a major limitation of many classical neuropsychological tests. Not least, this aspect also constitutes a prerequisite for applying the paradigm repeatedly, for instance in a pre- and post-

intervention study design. Second, the task is judged on how well its difficulty can be (parametrically) adjusted. The required levels of difficulty may vary markedly between study populations (e.g., patients vs. healthy controls, younger vs. older participants) or across multiple sessions in repeated within- participant application. Therefore, the task is evaluated with respect to experimental parameters that can be effectively manipulated to systematically affect task performance. Related, it should be consid- ered if task difficulty is adaptable enough to induce sufficient across-participant performance variance and avoid floor and ceiling effects. Since immersive VR-based paradigms are centered around user-interactions with the virtual environment in real-time (i.e., adapting the subjective view according to the individual’s position), they also embrace the possibility of highly dynamic testing scenarios, thus allowing the implementation of assessments in accordance with the ideas of computerized adaptive testing (van der Linden & Glas, 2000). It can be evaluated whether the task implements such an approach.

## Performance quantification

A further important prerequisite for a suitable candidate paradigm concerns the ability to measure user performance in a quantitative way. One should therefore consider if outcome variables to quantify per- formance have been defined, or if they can be derived from the data obtained in VR. As behavior in virtual environments can be tracked digitally with high resolution in both time and space, VR offers increasingly multivariate and experimenter-independent performance parameters, facilitating more ob- jective, data-driven and automated analysis approaches. It is therefore evaluated to what extent the candidate paradigm allows for experimenter-independent performance evaluation.

## Immersive capacities

Another dimension not adequately captured by traditional quality criteria concerns the capacity of VR systems to create the illusion of being located in the virtual world. There is an ongoing conceptual debate about the technical terms describing this phenomenon, specifically "immersion" and its relation to and disentanglement from the notion of "presence" [(Sanchez-Vives & Slater, 2005;](#_bookmark69) [Slater, 1999,](#_bookmark73) [2003;](#_bookmark74) [Witmer](#_bookmark84) [& Singer,](#_bookmark84) [1998).](#_bookmark84) For the purpose of paradigm development, we follow Slater in the distinction that immersion describes a VR system's objective technical properties that support natural sensorimotor contingencies, while presence refers to the subjective illusion of "being there" in the environment as a subjective correlate of immersion [(Slater,](#_bookmark75) [2009,](#_bookmark75) [2018;](#_bookmark76) [Slater & Sanchez-Vives,](#_bookmark77) [2016).](#_bookmark77) Accordingly, one first evaluates the degree of immersion as specified by task factors and the VR system necessary to present this task. Second, the likelihood that the task (in its final implementation) will facilitate the illusion of being in the virtual environment is considered, and ideally this judgment is informed by prior empirical evidence using presence questionnaires. This evaluation is important for two reasons: first, the degree to which participants feel present in the virtual environment may either have direct implications for the research question or represent a latent factor influencing task performance or user engagement, constituting a potential confound. Second, state-of-the-art VR technology raises hopes that

a higher degree of presence could be beneficial in diagnostic assessment, cognitive training outcome, or user experience (with respect to the latter, cf. [Brade et al.](#_bookmark12) [2017,](#_bookmark12) and [Lorenz et al.](#_bookmark47) [2018).](#_bookmark47) Indeed, there is some evidence that increased presence may have a positive impact on participants' cognitive performance, for instance regarding fact learning [(Cheng, She, & Annetta,](#_bookmark15) [2015)](#_bookmark15) or memory encoding [(Makowski, Sperduti, Nicolas, & Piolino,](#_bookmark51) [2017),](#_bookmark51) although potential benefits of increased presence in clinical assessment remain to be explored.

## Training feasibility

A further consideration concerns the feasibility of the candidate paradigm to serve as the basis for a training tool. In a one-time application setting (e.g., purely diagnostic assessment), the paradigm needs to fulfill fewer requirements compared to a repeated-application setting (e.g., implementation of a cognitive training tool). First, one should evaluate whether there are any practical obstacles to the repeated application of the paradigm. On the one hand, this concerns the logistics of the VR system used for the task: Can the task be administered in multiple sites or at home, or must the user be tested in a specialized laboratory, for instance due to the necessity of specific interaction devices such as a VR treadmill or a cave automatic virtual environment (CAVE, e.g., [Gromer et al.,](#_bookmark25) [2018)?](#_bookmark25) On the other hand, potential caveats in user feasibility may yield cumulative disadvantages in the training scenario (e.g., mild risk of cybersickness may be acceptable in a one-time application, but could decrease compliance when repeated with high frequency). Second, one determines if the necessary prerequisites of task adaptability are met (e.g., possibility to create parallel versions, effective manipulation of difficulty). Third, it is evaluated to what extent the paradigm offers the possibility of conveying cognitive strategies for compensatory training and how these could be implemented, for instance by leveraging the extensive cueing possibilities in VR [(Rizzo & Koenig,](#_bookmark66) [2017).](#_bookmark66) Furthermore, the likelihood of transfer effects is examined and if there is any empirical indication on their expected quality regarding near vs. far transfer.

## Predictable pitfalls

Furthermore, it is advisable to vet the candidate paradigm for predictable pitfalls. As in any clinical study, implementing a VR paradigm for cognitive assessment requires time, know-how, and monetary resources that must be weighed against potential knowledge gains and patient benefit. In order to optimize the potential of the research endeavor, one first evaluates how well the candidate paradigm adheres to the task requirements of the current research project and if the paradigm can be modified to maximize this adherence. Second, it is considered to what extent the application of the candidate

paradigm constitutes a reasonable allocation of the study resources. Not least, scrutinizing potential pitfalls early on in the development process also serves as a quality check when designing a VR paradigm de novo.

# Application of the VR-Check framework

The following illustrates how systematic evaluation with the VR-Check framework can guide the decision- making process in defining a neuropsychological VR paradigm for a specific research project.

## Evaluation workflow

First, the properties required of the VR paradigm are defined. Of note, these task requirements are nec- essarily project-specific, such that the relative weight of the various evaluation dimensions will naturally differ across projects. Furthermore, to facilitate a comparative evaluation of tasks across the VR-Check features, a semiquantitative rating is applied to evaluate if a particular feature applies to the candi- date task to a high, medium, or low degree, or if there is insufficient information to make an informed judgment. Researchers are thus able to systematically go through the list of features and judge each candidate paradigm according to the description above. For the de novo development of VR paradigms, the same process is applied to competing ideas or prototypes, yielding an explicit account of which task features need to be maximized.

## Example project: task requirements and candidate paradigms

Here, the results of this evaluation procedure are presented for an exemplary research project emanating from our consortium. It is important to note that the following outcomes do not represent a judgment on the value of the paradigms per se, but rather provide an illustration of the evaluation process itself and how it can inform project-specific paradigm optimization.

The goal of the exemplary research project was to apply immersive VR for the neuropsychological assessment of spatial cognition and executive functions. Suitable tasks were required to (i) be relevant to participants' everyday life, (ii) be feasible in a wide range of neurological patient populations, (iii) inform the development of a subsequent cognitive training tool, (iv) be implemented with an HMD, and (v) allow for natural user interaction based on body-tracking devices and gesture recognition. Candidate paradigms were identified by literature screening of existing VR tasks and in-house paradigms from January-May 2018. All candidate tasks were assessed along the VR-Check dimensions by an interdisciplinary research consortium including cognitive neuroscientists, physicians, and clinical neuropsychologists. Figure [2B](#_bookmark3)

visualizes the consensus ratings for a subset of four promising candidate tasks in each cognitive domain. While the below is limited to a brief account of the most decisive aspects, the interested reader is referred to the [Appendix](#_bookmark86) for a detailed point-by-point description of the systematic evaluation.

## Example project: spatial cognition

With respect to spatial cognition, the evaluation process is illustrated in the following candidate tasks:

1) the Starmaze (STM, [Igl6i, Doeller, Berthoz, Rondi-Reig, & Burgess,](#_bookmark31) [2010;](#_bookmark31) [Igl6i et al.,](#_bookmark32) [2014;](#_bookmark32) [Igl6i, Zaoui, Berthoz, & Rondi-Reig, 2009)](#_bookmark33) - a VR adaptation of a rodent paradigm [(Rondi-Reig et al., 2006)](#_bookmark68) to differentiate egocentric from allocentric navigation strategies - in which the user navigates through a point-symmetric star-shaped labyrinth in order to find a target; 2) the Virtual Memory Task (VMT, [Koenig et al.,](#_bookmark40) [2011),](#_bookmark40) a computerized spatial memory task similar to an existing real-life task [(Ploner, Stenz, Fassdorf, & Arnold, 2005),](#_bookmark62) in which participants are required to memorize locations of everyday objects on a table; 3) the Virtual Morris Water Maze (vMWM, e.g. [Korthauer, Nowak, Frahmand, & Driscoll, 2017;](#_bookmark41) [Nedelska et al., 2012),](#_bookmark57) a VR adaptation of the classical place navigation task originating from rodent research [(Morris, Garrud, Rawlins, & O'Keefe, 1982),](#_bookmark55) in which participants learn to navigate to a concealed platform; 4) the Cognitive Map Task (CMT, [Iaria, Chen, Guariglia, Ptito, & Petrides, 2007),](#_bookmark29) a spatial learning paradigm in which participants have to construe, maintain, and retrieve a cognitive map of a virtual town by learning and finding landmarks.

Based on the assessment along the VR-Check dimensions, the STM and the vMWM - while certainly highly appropriate paradigms for other research questions - were judged to be less favorable for our purposes due to limited ecological relevance, task adaptability, and training potential. In contrast, the VMT emerged as the paradigm that most closely adhered to our task requirements, made explicit through point-by-point assessment along the VR-Check dimensions: beside high ecological relevance to our target populations, favorable user feasibility and excellent adaptability, it further avoids some of the caveats of other candidate paradigms (such as high navigation complexity or the risk of adverse effects) and demands comparatively moderate implementation efforts, rendering it the optimal allocation of our study resources. Nonetheless, the VMT is limited to an assessment of spatial memory capacities due to the comparatively narrow domain target. In terms of assessing navigational abilities, the CMT was evaluated to be the most suitable starting point for the development of an immserive paradigm because of favorable ecological relevance, user feasibility and motivation, and high training potential. Notwithstanding, our evaluation process also identified potential improvements of the CMT that have to be addressed in the development process, such as a more fine-grained adaptation of difficulty.

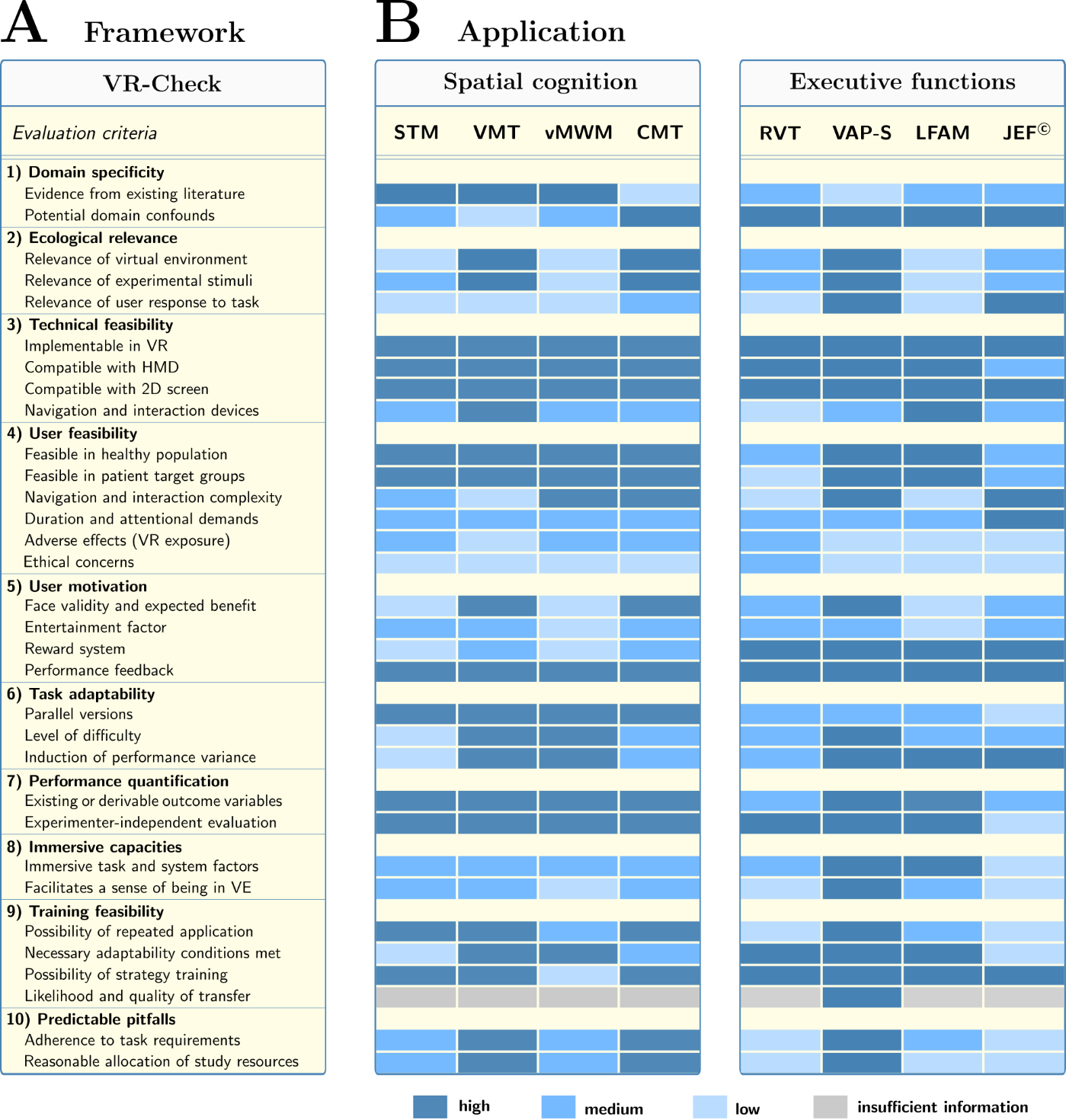


Figure 2: The VR-Check framework for Virtual Reality paradigms in neuropsychology. Panel A summarizes the evaluation dimensions in form of a checklist, as detailed in sec. [2.](#_bookmark1) Panel B visualizes the application of the framework for the exemplary cases of assessing spatial cognition and executive functions with immersive VR. The color schemes display the consensus ratings on whether the particular item applies to the candidate paradigm to a high, medium, or low degree. The evaluation procedure is illustrated here for the Starmaze (STM), Virtual Memory Task (VMT), the virtual Morris Water Maze (vMWM), and the Cognitive Map Test (CMT) for assessing spatial abilities, and the Ride in a Virtual Town (RVT), the Virtual Action Planning-Supermarket (VAP-S), the Look for a Match task (LFAM) as well as the Jansari Assessment of Executive Functions (JEF*Q*C ) for assessing executive functions. For the given task requirements, the VMT, CMT, and the VAP-S emerged as the most suitable paradigms for the development of an immersive VR application, as detailed in sec. [3.](#_bookmark2) HMD: head-mounted display; VE: virtual environment; VR: virtual reality.

## Example project: executive functions

Executive functions is an umbrella term for a multifaceted construct including several interconnected high-level cognitive abilities that serve ongoing, goal-directed actions [(Diamond,](#_bookmark19) [20](#_bookmark19)13). Subdomains include planning, problem solving, monitoring, working memory, inhibition, and task switching, and despite ongoing terminological disambiguations, there is relative agreement on the complexity and su- perordinate coordination role of executive functions and their importance regarding human adaptive behavior [(Diamond, 2013;](#_bookmark19) [Hofmann, Schmeichel, & Baddeley, 2012;](#_bookmark27) [Jurado & Rosselli, 2007;](#_bookmark37) [Martin &](#_bookmark52) [Failows, 2010).](#_bookmark52) For a comprehensive review of executive functions paradigms in VR see [Parsons (2015)](#_bookmark59) and [Valladares-Rodriguez, Perez-Rodriguez, Anido-Rif6n, and Fernandez-Iglesias (2016).](#_bookmark81) As above, we exemplify evaluation outcomes in four candidate paradigms: 1) A Ride in a Virtual Town (RVT, [Lecou-](#_bookmark45) [vey et al., 2017),](#_bookmark45) a prospective memory task featuring a car drive using real car components as interaction devices while completing a list of errands; 2) the Virtual Action Planning- Supermarket (VAP-S, [Klinger](#_bookmark38) [et al., 2006),](#_bookmark38) a grocery shopping task; 3) the Look for a Match task (LFAM, [Elkind, Rubin, Rosenthal,](#_bookmark21) [Skoff, & Prather, 2001),](#_bookmark21) an adaptation of the Wisconsin Card Sorting Task to a virtual beach environ-

ment; 4) the Jansari Assessment of Executive Functions (JEF*Q*C task requiring multitasking to prepare a meeting on time.

[Jansari et al., 2014),](#_bookmark34) a multi-step office

Resulting from the VR-Check evaluation, some inconsistencies with our task requirements were identi- fied for the LFAM (limited ecological relevance to our target populations, drawbacks in user motivation), the RVT (risk of adverse effects, incompatibilities with our interaction requirements, ecological relevance limited to drivers, ethical concerns about loss of driving capability in patient population, limited train-

ing feasibility) as well as the JEF*Q*C

(user feasibility limited to higher-functioning population, ecological

relevance restricted to a subgroup of our target population, incompatibilities with our immersive system factors, limited training feasibility due to caveats in task adaptability). The VAP-S, in contrast, was evaluated to be highly consistent with the project's task requirements regarding user feasibility, technical requirements, ecological relevance, and training potential, while demanding reasonable implementation efforts. The VAP-S was therefore esteemed the most favorable basis for the development of an immersive executive functions paradigm. Nonetheless, the systematic evaluation also highlighted potential caveats of the paradigm (limited domain specificity, technical solution required to multidirectional large-scale locomotion), which can thus be explicitly optimized in the implementation process.

# Discussion

In order to leverage the potential of VR in neuropsychology, researchers are increasingly challenged with optimizing the experimental paradigm to address the study question at hand. In light of the fast-growing body of literature and the increasing availability of VR hard- and software, the need arises for a new methodological framework on systematic paradigm evaluation. This gap is aggravated further by the inability of traditional psychometric quality criteria to capture the manifold properties of contemporary VR. With the present work, we aim to address this gap with a multidimensional evaluation protocol for VR applications in neuropsychology, summarized as an easy-to-use checklist (VR-Check).

## Paradigm optimization and across-domain tradeoffs

The systematic evaluation approach of the VR-Check framework raises the general question of what constitutes an ideal VR paradigm for neuropsychological research. Surely, if we defined an entertaining, highly adaptable, easy-to-play, easy-to-implement, highly immersive task that is viable for any user group, targets a well-circumscribed cognitive domain, adequately captures cognitive deficits relevant to everyday functioning as measured objectively by experimenter-independent performance outcomes, and which can be applied repeatedly to induce systematic improvement in both the tested and further cognitive domains - such a paradigm would be welcomed by researchers and clinicians alike.

As a corollary of the multidimensional nature of VR, however, such an endeavor is unrealistic for two principled reasons: first, what is desired of the task is tightly linked to the research question of interest. In consequence, there is no general profile of objectively desirable properties. While minimal requirements regarding user feasibility or technical implementation must be met by any clinical paradigm, the relative importance of the various domains will differ markedly over research applications and target populations. Indeed, the VR-Check framework serves precisely the purpose of prioritizing which domains are more important than others to address a given research question. This flexibility towards the study purpose enables researchers to weigh the different dimensions against each other and maximize the adherence to their project-specific requirements.

Second, the VR-Check framework illustrates a qualitative difference with respect to the interaction among evaluation criteria in that some are logically congruous, while others imply reciprocal incon- gruities. For instance, a paradigm featuring high training feasibility must also fulfill a variety of re- quirements concerning technical feasibility, user feasibility, and task adaptability, and is more likely to be judged favorably in terms of user motivation because these dimensions to some extent inform the evaluation of training feasibility. In contrast, other comparisons yield across-domain tradeoffs: Specif- ically, this concerns the relationship between cognitive domain specificity and ecological relevance. In

the attempt to target a specific cognitive domain with high precision, the recruitment of other cognitive domains must be minimized. However, this is rarely the case in everyday functioning, when a multitude of cognitive domains are engaged simultaneously. A VR paradigm featuring high ecological relevance will therefore necessarily concede some domain specificity by recruiting other domains than the one intended. Inversely, a VR paradigm featuring high domain specificity permits only limited relevance to cognitive functioning in real life because of an artificially narrow cognitive target. Due to this incongruousness, a deliberate decision must be made on the tradeoff between domain specificity and ecological relevance.

A similar point arises with respect to the relationship between ecological relevance and experimen- tal control. While both task and environment are highly controllable in VR, the increased degrees of behavioral freedom can result in less controlled participant behavior as compared to classical neuropsy- chological assessments. This behavioral freedom comes with an increased number of error sources not related to the cognitive task itself, such as visual attention, motor control, or navigational demands. In the research context, we can increase experimental control by restricting what the participant can and cannot do in VR. However, this again entails decreased relevance to everyday functioning, as real-life behavior offers similarly many degrees of freedom and also encompasses a multitude of error sources.

In sum, the properties required of a VR paradigm are dependent on the research question at hand, and there are inevitable across-dimension tradeoffs in paradigm design. These aspects necessitate deliberate design decisions to permit the project-specific optimization of the VR paradigm. The VR-Check frame- work guides this optimization process because it allows for a systematic account of how well a paradigm adheres to the project-specific requirements and because it makes these design decisions explicit.

## Towards improved standardization of clinical YR applications

While the assets of VR for clinical research have been examined before [(Birckhead et al.,](#_bookmark10) [2019;](#_bookmark10) [Lange](#_bookmark44) [et al., 2012;](#_bookmark44) [Rizzo et al., 2000;](#_bookmark64) [Rizzo & Kim, 2005;](#_bookmark65) [Rizzo & Koenig, 2017;](#_bookmark66) [Rizzo et al., 2004),](#_bookmark67) previous approaches have predominantly addressed general favorable properties of the technology, focused on specific therapeutic VR applications, or dealt with clinical study design for VR-based therapies. The VR-Check framework complements this prior work because it specifically targets the optimization of the paradigm (rather than the study) design, explicitly addresses cognitive and behavioral research, and because it provides researchers with a pragmatic easy-to-use evaluation tool. However, even though the application of the framework as described in section [3](#_bookmark2) was highly informative in the exemplary research project, some limitations of the present work deserve mentioning. First, application of the framework was limited to the assessment of spatial cognition and executive functions, such that further

research is necessary to corroborate its utility with respect to other research questions. Moreover, current evaluation outcomes are limited to semiquantitative assessment and consensus ratings, warranting further work to solidify more quantitative approaches. Furthermore, it should be noted that paradigms that have been applied more often in the literature might lead researchers to evaluate them more favorably simply because existing evidence makes these paradigms easier to judge. However, the fact that a paradigm may be more established in the literature does not necessarily imply that it is better suited for the study question at hand. Finally, we here focused on the design optimization of VR paradigms for neuropsychological assessment. Nonetheless, the value of therapeutic VR applications is becoming increasingly apparent [(Birckhead et al., 2019;](#_bookmark10) [Coyle, Traynor, & Solowij, 2015;](#_bookmark16) [Garcia-Betances, Jimenez-](#_bookmark23) [Mixco, Arredondo, & Cabrera-Umpierrez, 2015),](#_bookmark23) and there is an important interplay between assessment and rehabilitation, especially with regard to devising individualized therapies that cater to the patient's specific deficits ("precision medicine"). While many of the VR-Check dimensions appear relevant to clinical VR tasks in general (e.g., technical and user feasibility, adaptability, or outcome quantification), future work must investigate if the protocol is also applicable to VR tools for cognitive training and rehabilitation, or to what extent the framework must be modified to enable paradigm optimization for these applications.

Even with these limitations in mind, the VR-Check framework represents a first step towards stan- dardized optimization of VR paradigms in clinical neuropsychology. The potential of contemporary VR is contrasted by a relative scarcity of consensus on what should be regarded as best practice when applying VR in clinical research. With respect to paradigm development, the VR-Check framework can inform this discussion. Even with optimal paradigm design, however, proof of clinical utility ul- timately requires high-quality empirical evidence such as randomized controlled trials (RCTs). In this context, the newly established Virtual Reality Committee of Outcomes Research Experts (VR-CORE) has recently suggested a framework for the development and validation of VR-based therapies [(Birck-](#_bookmark10) [head et al.,](#_bookmark10) [2019).](#_bookmark10) This framework features three study phases (VR1-VR3) similar to the phase I-III model of pharmacological intervention studies. While the authors' approach specifically focuses on VR treatments, their systematic methodological framework is similar in spirit to our suggestions, and the two approaches complement each other (paradigm design optimization and study design optimization). For instance, the authors' call for human-centered design in early VR treatment study phases (VR1) is matched by our explicit focus on the patient perspective in the domains of technical feasibility, user fea- sibility, user motivation, and ecological relevance. The intermediate trial phase (VR2) is concerned with initial feasibility testing and can thus be regarded as the study design counterpart to the paradigm design feasibility dimensions of the VR-Check framework. The later VR-CORE phase (VR3) concerns RCTs to examine VR treatment efficacy versus a control condition. As such, the extension of the VR3 phase

to neuropsychological assessment tasks seems natural: where a VR treatment must show intervention efficacy, a VR assessment task must show discriminatory or predictive power in empirical evaluation.

As methodological guidelines such as the VR-CORE recommendations and the VR-Check framework are further developed, they may ultimately synergize in pursuit of a more rigorous, systematic, and well-informed protocol for the development of clinical VR applications.

# Acknowledgments

The present work emanates from the research consortium VReha - virtual worlds for digital diagnostics and cognitive rehabilitation, funded by the German Federal Ministry for Education and Research (BMBF grant 13GW0206).

# References

Astur, R. S., Taylor, L. B., Mamelak, A. N., Philpott, L., & Sutherland, R. J. (2002). Humans with hippocampus damage display severe spatial memory impairments in a virtual Morris water task. Behavioural brain research, 132 (1), 77-84.

Astur, R. S., Tropp, J., Sava, S., Constable, R. T., & Markus, E. J. (2004). Sex differences and correla- tions in a virtual Morris water task, a virtual radial arm maze, and mental rotation. Behavioural brain research, 151 (1-2), 103-115.

Aubin, G., Beliveau, M.-F., & Klinger, E. (2018). An exploration of the ecological validity of the Virtual Action Planning-Supermarket (VAP-S) with people with schizophrenia. Neuropsychological rehabilitation, 28 (5), 689-708.

Aukstakalnis, S., & Blatner, D. (1992). Silicon mirage; the art and science of virtual reality. Peachpit Press.

Bellassen, V., Igl6i, K., de Souza, L. C., Dubois, B., & Rondi-Reig, L. (2012). Temporal order mem- ory assessed during spatiotemporal navigation as a behavioral cognitive marker for differential Alzheimer's disease diagnosis. Journal of Neuroscience, 32 (6), 1942-1952.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. The Journal of general psychology , 39 (1), 15-22.

Birckhead, B., Khalil, C., Liu, X., Conovitz, S., Rizzo, A., Danovitch, I., . . . Spiegel, B. (2019, Jan 31). Recommendations for methodology of virtual reality clinical trials in health care by an international working group: Iterative study. JMIR Ment Health, 6 (1), e11973. Retrieved from [https://](https://mental.jmir.org/2019/1/e11973/) [mental.jmir.org/2019/1/e11973/](https://mental.jmir.org/2019/1/e11973/) doi: 10.2196/11973

Bornstein, R. F. (1996). Face validity in psychological assessment: Implications for a unified model of validity.

Brade, J., Lorenz, M., Busch, M., Hammer, N., Tscheligi, M., & Klimant, P. (2017). Being there again- presence in real and virtual environments and its relation to usability and user experience using a mobile navigation task. International Journal of Human-Computer Studies, 101 , 76-87.

Brookes, J., Warburton, M., Alghadier, M., Mon-Williams, M., & Mushtaq, F. (2019). Studying human behavior with virtual reality: The unity experiment framework. Behavior research methods, 1-9.

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.

Cheng, M.-T., She, H.-C., & Annetta, L. A. (2015). Game immersion experience: its hierarchical structure and impact on game-based science learning. Journal of Computer Assisted Learning ,  31 (3), 232-253.

Coyle, H., Traynor, V., & Solowij, N. (2015). Computerized and virtual reality cognitive training for individuals at high risk of cognitive decline: systematic review of the literature. The American Journal of Geriatric Psychiatry, 23 (4), 335-359.

Dawson, D. R., & Marcotte, T. D. (2017). Special issue on ecological validity and cognitive assessment.

Taylor & Francis.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gameful- ness: defining gamification. In Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments (pp. 9-15).

Diamond, A. (2013). Executive functions. Annual review of psychology , 64 , 135-168.

Eling, P., Derckx, K., & Maes, R. (2008). On the historical and conceptual background of the Wisconsin Card Sorting Test. Brain and cognition, 67 (3), 247-253.

Elkind, J. S., Rubin, E., Rosenthal, S., Skoff, B., & Prather, P. (2001). A simulated reality scenario compared with the computerized Wisconsin Card Sorting Test: An analysis of preliminary results. CyberPsychology 8 Behavior, 4 (4), 489-496.

Franzen, M. D., & Wilhelm, K. L. (1996). Conceptual foundations of ecological validity in neuropsycho- logical assessment.

Garcia-Betances, R. I., Jimenez-Mixco, V., Arredondo, M. T., & Cabrera-Umpierrez, M. F. (2015). Using virtual reality for cognitive training of the elderly. American Journal of Alzheimer s Disease 8 Other Dementias*o*R , 30 (1), 49-54.

Gilboa, Y., Jansari, A., Kerrouche, B., Ucak, E., Tiberghien, A., Benkhaled, O., . . . Abada, G. (2017). 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 Func-

tions for Children (JEF-C*o*c ). Neuropsychological rehabilitation, 1-24.

Gromer, D., Madeira, O., Gast, P., Nehfischer, M., Jost, M., Miiller, M., . . . Pauli, P. (2018). Height simulation in a virtual reality cave system: validity of fear responses and effects of an immersion manipulation. Frontiers in human neuroscience, 12 , 372.

Griibel, J., Weibel, R., Jiang, M. H., Holscher, C., Hackman, D. A., & Schinazi, V. R. (2016). EVE: a framework for experiments in virtual environments. In Spatial cognition x (pp. 159-176). Springer. Hofmann, W., Schmeichel, B. J., & Baddeley, A. D. (2012). Executive functions and self-regulation.

Trends in cognitive sciences, 16 (3), 174-180.

Howard, M. C. (2018). Virtual reality interventions for personal development: A meta-analysis of hardware and software. Human-Computer Interaction, 1-35.

Iaria, G., Chen, J.-K., Guariglia, C., Ptito, A., & Petrides, M. (2007). Retrosplenial and hippocampal brain regions in human navigation: complementary functional contributions to the formation and use of cognitive maps. European Journal of Neuroscience, 25 (3), 890-899.

Iaria, G., Palermo, L., Committeri, G., & Barton, J. J. (2009). Age differences in the formation and use of cognitive maps. Behavioural brain research, 196 (2), 187-191.

Igl6i, K., Doeller, C. F., Berthoz, A., Rondi-Reig, L., & Burgess, N. (2010). Lateralized human hip- pocampal activity predicts navigation based on sequence or place memory. Proceedings of the National Academy of Sciences, 107 (32), 14466-14471.

Igl6i, K., Doeller, C. F., Paradis, A.-L., Benchenane, K., Berthoz, A., Burgess, N., & Rondi-Reig, L. (2014). Interaction between hippocampus and cerebellum crus i in sequence-based but not place- based navigation. Cerebral Cortex , 25 (11), 4146-4154.

Igl6i, K., Zaoui, M., Berthoz, A., & Rondi-Reig, L. (2009). Sequential egocentric strategy is acquired as early as allocentric strategy: Parallel acquisition of these two navigation strategies. Hippocampus,  19 (12), 1199-1211.

Jansari, A. S., Devlin, A., Agnew, R., Akesson, K., Murphy, L., & Leadbetter, T. (2014). Ecological assessment of executive functions: a new virtual reality paradigm. Brain Impairment, 15 (2), 71-87. Josman, N., Kizony, R., Hof, E., Goldenberg, K., Weiss, P. L., & Klinger, E. (2014). Using the virtual action planning-supermarket for evaluating executive functions in people with stroke. Journal of

Stroke and Cerebrovascular Diseases, 23 (5), 879-887.

Josman, N., Klinger, E., & Kizony, R. (2008). Performance within the virtual action planning super- market (VAP-S): an executive function profile of three different populations suffering from deficits in the central nervous system. Maia, Portugal: Proc 7th ICDVRAT.

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: a review of our current understanding. Neuropsychology review, 17 (3), 213-233.

Kizony, R., Korman, M., Sinoff, G., Klinger, E., Josman, N., & Sharky, P. (2012). Using a virtual supermarket as a tool for training executive functions in people with mild cognitive impairment. In Proceedings of the 9th international conference on disability, virtual reality and associated tech- nologies (pp. 41-46).

Klinger, E., Chemin, I., Lebreton, S., & Marie, R.-M. (2006). Virtual action planning in Parkinson's disease: A control study. Cyberpsychology 8 Behavior, 9 (3), 342-347.

Koenig, S., Crucian, G. P., Diinser, A., Bartneck, C., & Dalrymple-Alford, J. C. (2011). Validity evaluation of a spatial memory task in virtual environments. Int J Des Innov Res, 6 , 1-13.

Korthauer, L., Nowak, N., Frahmand, M., & Driscoll, I. (2017). Cognitive correlates of spatial navigation: associations between executive functioning and the virtual morris water task. Behavioural brain research, 317 , 470-478.

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

Lalonde, G., Henry, M., Drouin-Germain, A., Nolin, P., & Beauchamp, M. H. (2013). Assessment of executive function in adolescence: A comparison of traditional and virtual reality tools. Journal of Neuroscience Methods, 219 (1), 76-82.

Lange, B., Koenig, S., Chang, C.-Y., McConnell, E., Suma, E., Bolas, M., & Rizzo, A. (2012). Designing informed game-based rehabilitation tasks leveraging advances in virtual reality. Disability and rehabilitation, 34 (22), 1863-1870.

Lecouvey, G., Gonneaud, J., Piolino, P., Madeleine, S., Orriols, E., Fleury, P., . . . Desgranges, B. (2017). Is binding decline the main source of the ageing effect on prospective memory? A ride in a virtual town. Socioafective neuroscience 8 psychology , 7 (1), 1304610.

Liu, I., Levy, R. M., Barton, J. J., & Iaria, G. (2011). Age and gender differences in various topographical orientation strategies. Brain research, 1410 , 112-119.

Lorenz, M., Brade, J., Diamond, L., Sjolie, D., Busch, M., Tscheligi, M., . . . Hammer, N. (2018). Presence and user experience in a virtual environment under the influence of ethanol: An explorative study. Scientific reports, 8 (1), 6407.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafo, M. R. (2016). Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. JMIR serious games, 4 (2), e11.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: an integrative review. Psycho- logical bulletin, 109 (2), 163.

Madary, M., & Metzinger, T. K. (2016). Real virtuality: a code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. Frontiers in Robotics and AI , 3 ,  3.

Makowski, D., Sperduti, M., Nicolas, S., & Piolino, P. (2017). "Being there" and remembering it: Presence improves memory encoding. Consciousness and cognition, 53 , 194-202.

Martin, J., & Failows, L. (2010). Executive function: Theoretical concerns. In J. Fagerberg, D. C. Mow- ery, & R. R. Nelson (Eds.), Self and social regulation - social interaction and the development of social understanding and executive functions (chap. 2). Oxford, GB: Oxford University Press.

Moffat, S. D., & Resnick, S. M. (2002). Effects of age on virtual environment place navigation and allocentric cognitive mapping. Behavioral neuroscience, 116 (5), 851.

Moosbrugger, H., & Kelava, A. (2012). Testtheorie und Fragebogenkonstruktion. Springer.

Morris, R., Garrud, P., Rawlins, J. a., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. Nature, 297 (5868), 681.

Navarro, M.-D., Llorens, R., Noe, E., Ferri, J., & Alcafiiz, M. (2013). Validation of a low-cost virtual reality system for training street-crossing. a comparative study in healthy, neglected and non- neglected stroke individuals. Neuropsychological rehabilitation, 23 (4), 597-618.

Nedelska, Z., Andel, R., Lacz6, J., Vlcek, K., Horinek, D., Lisy, J., . . . Hort, J. (2012). Spatial navigation impairment is proportional to right hippocampal volume. Proceedings of the National Academy of Sciences, 109 (7), 2590-2594.

Negut, 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.

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 , 660.

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 8 Behavior, 11 (1), 17-25.

Pedroli, E., Serino, S., Giglioli, A. C., Pallavicini, F., Cipresso, P., & Riva, G. (2016). The use of virtual reality tools for the assessment of executive functions and unilateral spatial neglect. In Virtual reality enhanced robotic systems for disability rehabilitation (pp. 115-140). IGI Global.

Ploner, C. J., Stenz, U., Fassdorf, K., & Arnold, G. (2005). Egocentric and allocentric spatial memory in idiopathic cervical dystonia. Neurology , 64 (10), 1733-1738.

Rizzo, 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.

Rizzo, A., Buckwalter, J. G., van der Zaag, C., Neumann, U., Thiebaux, M., Chua, C., . . . Larson,

P. (2000). Virtual environment applications in clinical neuropsychology. In Virtual reality, 2000. proceedings. ieee (pp. 63-70).

Rizzo, A., & Kim, G. J. (2005). A SWOT analysis of the field of virtual reality rehabilitation and therapy. Presence: Teleoperators 8 Virtual Environments, 14 (2), 119-146.

Rizzo, A., & Koenig, S. T. (2017). Is clinical virtual reality ready for primetime? Neuropsychology ,  31 (8), 877.

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

Rondi-Reig, L., Petit, G. H., Tobin, C., Tonegawa, S., Mariani, J., & Berthoz, A. (2006). Impaired sequential egocentric and allocentric memories in forebrain-specific-nmda receptor knock-out mice during a new task dissociating strategies of navigation. Journal of Neuroscience, 26 (15), 4071-  4081.

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality.

Nature Reviews Neuroscience, 6 (4), 332.

Schultheis, M. T., Himelstein, J., & Rizzo, A. A. (2002). Virtual reality and neuropsychology: upgrading the current tools. The Journal of head trauma rehabilitation, 17 (5), 378-394.

Schultheis, M. T., & Rizzo, A. A. (2001). The application of virtual reality technology in rehabilitation.

Rehabilitation psychology , 46 (3), 296.

Shallice, T., & Burgess, P. W. (1991). Deficits in strategy application following frontal lobe damage in man. Brain, 114 (2), 727-741.

Slater, M. (1999). Measuring presence: A response to the Witmer and Singer presence questionnaire.

Presence, 8 (5), 560-565.

Slater, M. (2003). A note on presence terminology. Presence connect, 3 (3), 1-5.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical Transactions of the Royal Society B: Biological Sciences, 364 (1535), 3549-3557.

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. British journal of psychology (London, England: 1953), 109 (3), 431.

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. Frontiers in Robotics and AI , 3 , 74.

Spooner, D. M., & Pachana, N. A. (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.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of experimental psychology ,

18 (6), 643.

Teel, E., Gay, M., Johnson, B., & Slobounov, S. (2016). Determining sensitivity/specificity of virtual reality-based neuropsychological tool for detecting residual abnormalities following sport-related concussion. Neuropsychology , 30 (4), 474.

Valladares-Rodriguez, S., Perez-Rodriguez, R., Anido-Rif6n, L., & Fernandez-Iglesias, M. (2016). Trends on the application of serious games to neuropsychological evaluation: a scoping review. Journal of biomedical informatics, 64 , 296-319.

Werner, P., Rabinowitz, S., Klinger, E., Korczyn, A. D., & Josman, N. (2009). Use of the virtual action planning supermarket for the diagnosis of mild cognitive impairment. Dementia and geriatric cognitive disorders, 27 (4), 301-309.

Wiener, J. M., Carroll, D., Moeller, S., Bibi, I., Ivanova, D., Allen, P., & Wolbers, T. (2019). A novel virtual-reality-based route-learning test suite: Assessing the effects of cognitive aging on navigation. Behavior Research Methods, 1-11.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. Presence, 7 (3), 225-240.

Zhao, H., Thrash, T., Wehrli, S., Holscher, C., Kapadia, M., Griibel, J., . . . Schinazi, V. R. (2018). A networked desktop virtual reality setup for decision science and navigation experiments with multiple participants. JoVE (Journal of Visualized Experiments)(138), e58155.

# Appendix

The following provides a detailed description of the evaluation process illustrated in section [3,](#_bookmark2) based on a systematic point-by-point application of the VR-Check dimensions.

## Example project: spatial cognition

Evaluations concerns the Starmaze (STM, [Igl6i et al., 2010,](#_bookmark31) [2014,](#_bookmark32) [2009),](#_bookmark33) the Virtual Memory Task (VMT, [Koenig et al., 2011),](#_bookmark40) the Virtual Morris Water Maze (vMWM, e.g. [Korthauer et al., 2017;](#_bookmark41) [Nedelska et](#_bookmark57) [al., 2012),](#_bookmark57) and the Cognitive Map Task (CMT, [Iaria et al](#_bookmark29)., [2007),](#_bookmark29) as detailed in the main text.

Domain specificity was rated high for the STM and vMWM (spatial memory and navigation), with moderate domain confounds of selective attention (STM) and executive capabilities (vMWM, [Korthauer](#_bookmark41) [et al., 2017).](#_bookmark41) The VMT was also evaluated to feature high domain specificity, as it closely targets visu- ospatial memory (without navigational demands) and showed congruency with construct-driven tests in prior empirical studies [(Koenig et al., 2011),](#_bookmark40) while potential domain confounds were judged to be compar- atively low (moderate short-term memory capacity and low motor demands). The CMT was evaluated to yield low domain specificity and high potential for domain confounds, as the task is dependent on strategy learning and action planning besides spatial memory and navigation. Ecological relevance of the virtual environment was rated low for the STM and the vMWM due to the labyrinth setting, while the VMT (house setting) and the CMT (town setting) were rated high in this aspect. Ecological relevance of the experimental stimuli was deemed high for the VMT (household objects) and the CMT (streets and shops), moderate for the STM (artificial maze but with natural landmarks such as mountains), and low for the vMWM (artificial pool). In contrast, ecological relevance of the user response to solve the task was considered low for the STM and vMWM (finding a target in a maze) as well as the VMT (relocating objects after removal) and moderate for the CMT (building a cognitive map for orientation). All tasks were regarded as highly VR-feasible since computerized versions already exist. However, tech- nical feasibility with respect to moving in the virtual environment was deemed moderate for the STM, vMWM, and CMT, as all feature large-scale locomotion (or alternatively swimming for the vMWM), increasing the difficulty of technical implementation with respect to tracking-based navigation. In this regard, feasibility was judged high for the VMT (hand-object interaction, low interaction complexity, no large-scale locomotion required). Feasibility in both healthy and patient populations were rated high for all of the examined paradigms [(Astur, Taylor, Mamelak, Philpott, & Sutherland,](#_bookmark4) [2002;](#_bookmark4) [Bellassen,](#_bookmark8) [Igl6i, de Souza, Dubois, & Rondi-Reig,](#_bookmark8) [2012;](#_bookmark8) [Iaria, Palermo, Committeri, & Barton,](#_bookmark30) [2009;](#_bookmark30) [Koenig et](#_bookmark40) [al.,](#_bookmark40) [2011;](#_bookmark40) [Liu, Levy, Barton, & Iaria,](#_bookmark46) [2011).](#_bookmark46) Navigation complexity was considered high for the CMT and vMWM (unrestricted multidirectional motion), moderate for the STM (multidirectional motion but

restricted by maze walls), and low for the VMT (no large-scale motion). While adverse effects were not a problem for any of the tasks in the 2D version, the potential for VR-induced kinetosis was deemed moderate for the STM, vMWM, and CMT due to the necessity of large-scale motion in the immersive setting, while the risk of adverse effects was judged to be low for the VMT. All paradigms were judged to be ethically innocuous under the condition that potential adverse effects could be kept to a minimum. In this context, it is noteworthy that [Koenig et al.](#_bookmark40) [(2011)](#_bookmark40) reported a few cases of negative emotional reactions when patients suddenly realized deficits in spatial memory upon conducting the 2D VMT.

Regarding user motivation, expected benefit was considered to be low for the STM and vMWM due to the artificial nature of the tasks, while it was deemed high for the VTM and CMT, as our target study population would likely be able to relate to the purpose of the assessment. The entertainment factor was judged moderate for the STM, VMT, and CMT in their current versions and low for the vMWM due to a comparatively higher degree of repetitiveness. Feasibility of a reward system was deemed low in the STM (as the task's focus is to disentangle navigational strategies rather than assess navigational abilities) and the vMWM (due to the repetitive character) and moderate for the VMT (e.g., new or surprise objects) and CMT (e.g., access a new part of town). The possibility of performance feedback was considered high for all tasks. With respect to task adaptability, the feasibility of creating parallel versions was judged to be high for all tasks. The possibility of effectively manipulating difficulty levels, however, was judged to be high for the vMWM (e.g., size of pool, size of platform) and the VMT (e.g., number of memory items, encoding intervals), moderate for the STM (e.g., number of landmarks or more alleys, but this could induce navigation strategy confounds), and low for the CMT (e.g., size of town or complexity of cognitive map, but this could increase executive demands, aggravating domain confounds). As regards the induction of performance variance, the STM has shown ceiling effects in young healthy participants and floor effects in patients with Alzheimer's Disease and Mild Cognitive Impairment [(Bellassen et al.,](#_bookmark8) [2012).](#_bookmark8) For the CMT, while there was some indication of ceiling effects in young healthy adults [(Iaria et](#_bookmark29) [al., 2007),](#_bookmark29) it also induced variance with age [(Iaria et al., 2009).](#_bookmark30) In contrast, performance variability was evaluated high for the vMWM (e.g., [Astur, Tropp, Sava, Constable, & Markus, 2004;](#_bookmark5) [Moffat & Resnick,](#_bookmark53) [2002)](#_bookmark53) as well as the VMT [(Koenig et al., 2011).](#_bookmark40)

With respect to performance quantification, objective outcome measures were abundant for the STM (e.g., number of correct trials, path lengths, trial times), the VMT (e.g., distance errors, accuracy by memory items, placement analysis, trial times), the vMWM (e.g., latency and distance to platform, path complexity, duration), and the CMT (e.g., mean time delay from optimal route, detour errors) alike. Experimenter-independent automated quantification of these measures was considered easily possible for all paradigms. All tasks were judged to be compatible with our immersive system requirements, although moderate difficulties were anticipated in implementing locomotion (STM, vMWM, and CMT) and hand-

object interaction (VMT), respectively. As gesture-based interaction was one of the requirements, the capacity to facilitate a sense of being in the virtual environment was considered medium for the STM (locomotion), VMT (object manipulation), and CMT (locomotion), and low for the vMWM (swimming). With regard to training feasibility, no general obstacles to repeated application were identified for any of the tasks, although some moderate drawbacks were identified for the vMWM due to the limited user motivation. While the STM was evaluated to show high potential for strategy training (as the paradigm disentangles between navigational strategies with high precision), there were concerns in how adaptable the task would be for training, which was also the case for the CMT, but to a lesser extent. The VMT was judged to fulfill both these properties to a high extent, while the vMWM was considered to possess low potential for strategy training. Likelihood of transfer was difficult to judge here, as none of the examined tasks has been developed into a training paradigm yet.

Inconsistencies with the defined task requirements were identified for the STM and the vMWM due to limited ecological relevance, task adaptability, and training potential. In contrast, the VMT closely adhered to our task requirements based on high ecological relevance to our target populations, favorable user feasibility and excellent adaptability, while the CMT was evaluated to be the most suitable starting point for the development of an immserive navigation paradigm because of favorable ecological relevance, user feasibility and motivation, and high training potential.

## Example project: executive functions

Evaluation outcomes are provided for A Ride in a Virtual Town (RVT, [Lecouvey et al., 2017),](#_bookmark45) the Virtual Action Planning- Supermarket (VAP-S, [Klinger et al., 2006),](#_bookmark38) the Look for a Match task (LFAM, [Elkind](#_bookmark21)

[et al., 2001),](#_bookmark21) and the Jansari Assessment of Executive Functions (JEF*Q*C in the main text.

[Jansari et al., 2014),](#_bookmark34) as outlined

Regarding domain specificity, we found that the RVT, LFAM and JEF*Q*C

all focus on moderately

narrow executive subdomains, such as prospective memory or mental flexibility. In contrast, the VAP-S lists several subdomains, all subserving the activity of daily living "shopping," rendering the targeted domain spectrum comparatively wider. As mentioned above, the nature of executive functions lies in their superordinate functions, such that overlap with subordinate cognitive domains such as attention or memory is to some extent inevitable. Accordingly, all four tasks were considered to feature high potential for domain confounds. The four paradigms were evaluated to show considerable differences in their ecological relevance to our target population: Shopping in a supermarket is a frequent everyday activity for most people including the elderly and many neurological patients, such that the VAP-S was considered to be highly ecologically relevant in terms of the environment, stimuli, and user response to the

task (cf. [Aubin, Beliveau, & Klinger, 2018).](#_bookmark6) In comparison, the ecological relevance of the environment and the stimuli in the RVT is restricted to drivers, and the user response to the task contains some artificial elements (e.g., oral execution of errands, town with only one street). Similarly, the ecological

relevance of the environment and the experimental stimuli in the JEF*Q*C

(office workspace) is limited to

white-collar workers, although the user response to the task (multitasking to meet a temporal deadline) was considered to be of high ecological relevance to the target population. In contrast, ecological relevance was rated low in the LFAM for the environment (beach bar), stimuli (set of umbrellas), and user response (find out the hidden rules about which customer has ordered which product). All tasks were evaluated to be feasible for 2D implementation, but there were concerns about the HMD compatibility of the

JEF*Q*C

due to the dependence on hard copy material for execution. Compatibility with our technical

interaction requirements (body-tracking and gesture recognition) was deemed high for the LFAM (hand-

object interaction), medium for the JEF*Q*C

and VAP-S (large-scale movement as locomotion), and low

for the RVT (large-scale movement in car while seated). Interaction complexity, on the other hand, was low for the RVT (unidirectional movement, oral commands) and the LFAM (no locomotion, manual

interaction), while it was considered high for the VAP-S and JEF*Q*C

(multidirectional large-scale motion

and object interaction). While the LFAM and VAP-S were considered highly feasible in both healthy and patient target populations (for the VAP-S, see [Josman et al. 2014;](#_bookmark35) [Josman, Klinger, and Kizony 2](#_bookmark36)008),

the JEF*Q*C

was judged feasible only in higher-functioning populations (although alternative scenarios to

overcome this limitation are being developed; [Gilboa et al.](#_bookmark24), [2017).](#_bookmark24) Similarly, the RVT was considered to show moderate and low feasibility in the target healthy and patient groups, respectively, due to the driving setting. Moreover, among the considered paradigms, the RVT is the only one to report significant drop-outs due to cybersickness (15% in an older population, [Lecouvey et al., 2017).](#_bookmark45) Furthermore, as a non-negligible amount of our patient group could be expected to have lost their license or capacities to drive, we judged that confronting patients with this often highly emotional topic raises ethical concerns.

Attentional demands were deemed high for the JEF*Q*C density) and moderate for the remaining tasks.

(due to the higher multitasking load and temporal

The expected benefit of the LFAM was evaluated to be low in our target populations due to the low ecological relevance, and similarly for the JEF*Q*C , as the majority of our patients would not be expected to relate to the office setting. The expected benefit of the RVT was considered medium (based on the errands-running part), while the VAP-S scored high due to the very common task scenario. The entertainment factor was deemed low for the LFAM due to the repetitive character (although it does compare well against the original WCST), and otherwise considered moderate. The VAP-S, however, is the only paradigm for which explicit patient-reported outcomes are available, showing medium to high enjoyment and satisfaction. All tasks were evaluated to lend themselves well to reward systems

and performance feedback. Concerning the creation of parallel versions, implementation demand differs among the paradigms. For example, producing a parallel large-scale town (RVT) requires higher VR modeling resources than rearranging shelves and products in a supermarket (VAP-S). Nonetheless, for both tasks, creating parallel versions based on varying errand lists poses comparably little implementation effort (and a similar adaptation would be possible for the LFAM). In contrast, creating parallel tasks

for each of the nine subdomains of the JEF*Q*C

would require considerably more development resources,

which was considered less feasible for our purposes. The grading of difficulty was deemed to be easy for the VAP-S (e.g., size of supermarket, complexity of shopping list), while the narrow task demands (LFAM), the reduction to list-solving (RVT; an action is performed by saying the action), and the implementational challenges (JEF*Q*C ) were judged to pose moderate obstacles to difficulty grading for the other paradigms, respectively. Induction of across-participant variance has been shown for all tasks, with the limitation of moderate ceiling effects in healthy young participants for the RVT. In terms of performance quantification, the VAP-S and LFAM both offer a range of time- and accuracy-based

outcome variables, whereas the RVT and the JEF*Q*C

in their current forms report a comparatively limited

number of (discretized) outcomes. The JEF*Q*C

requires time-consuming and non-automatized post-hoc

ratings of the performance, limiting the feasibility of experimenter-independent evaluation (although inter-rater reliability was high in a small pilot study), while this was esteemed well feasible for the remaining tasks.

As regards immersive capacities, compatibility issues with the required system factors were identified

for the JEF*Q*C

(HMD feasibility) and the RVT (driving-wheel interface but oral actions), while the VAP-S

and LFAM both lend themselves well to the gesture-recognition and tracking prerequisites. Likelihood of

presence was esteemed low for the JEF*Q*C

(limited to 2D presentation) and RVT (restricted natural inter-

action) and moderate for the LFAM (artificial beach environment in current form), while this aspect has been empirically assessed for the VAP-S, yielding good presence ratings [(Werner, Rabinowitz, Klinger,](#_bookmark82) [Korczyn, & Josman,](#_bookmark82) [2009).](#_bookmark82) With respect to training potential, repeated application was considered well feasible for the VAP-S, but limited for the RVT (risk of cybersickness), the LFAM (drawbacks in

user motivation due to repetitiveness), and the JEF*Q*C

(resource-demanding test setting). Furthermore,

some concerns about task adaptability were raised regarding the latter (lack of parallel versions), even though implementing strategy cues was deemed well feasible for all tasks. However, the VAP-S is the only paradigm that has already been used as a training task in a pilot study [(Kizony et al., 2012),](#_bookmark39) with promising results concerning training transfer.

Inconsistencies with our task requirements were identified for the LFAM (limited ecological relevance to our target populations, drawbacks in user motivation), the RVT (risk of adverse effects, incompati- bilities with our interaction requirements, ecological relevance limited to drivers, ethical concerns about

loss of driving capability in patient population, limited training feasibility) as well as the JEF*Q*C

(user

feasibility limited to higher-functioning population, ecological relevance restricted to a subgroup of our target population, incompatibilities with our immersive system factors, limited training feasibility due to caveats in task adaptability). The VAP-S, in contrast, was evaluated to be highly consistent with the project's task requirements regarding user feasibility, technical requirements, ecological relevance, and training potential, while demanding reasonable implementation efforts.

[View publication stats](https://www.researchgate.net/publication/337758605)