Plotnik *et al. J NeuroEngineering Rehabil (2021) 18:82*
https://doi.org/10.1186/s12984-021-00849-9

Journal of NeuroEngineering
and Rehabilitation

RESEARCH

Open Access

Multimodal immersive trail making-virtual reality paradigm to study cognitive-motor interactions

Meir Plotnik[1,2,3*], Oran Ben-Gal[1†], Glen M. Doniger[1,4†], Amihai Gottlieb[1], Yotam Bahat[1], Maya Cohen[1], Shani Kimel-Naor[1], Gabi Zeilig[5,6] and Michal Schnaider Beeri[4,7]

Abstract

Background: Neuropsychological tests of executive function have limited real-world predictive and functional relevance. An emerging solution for this limitation is to adapt the tests for implementation in virtual reality (VR). We thus developed two VR-based versions of the classic Color-Trails Test (CTT), a well-validated pencil-and-paper executive function test assessing sustained (Trails A) and divided (Trails B) attention—one for a large-scale VR system (DOME-CTT) and the other for a portable head-mount display VR system (HMD-CTT). We then evaluated construct validity, test–retest reliability, and age-related discriminant validity of the VR-based versions and explored effects on motor function.

Methods: Healthy adults ($n = 147$) in three age groups (young: $n = 50$; middle-aged: $n = 80$; older: $n = 17$) participated. All participants were administered the original CTT, some completing the DOME-CTT (14 young, 29 middle-aged) and the rest completing the HMD-CTT. Primary outcomes were Trails A and B completion times (t_A, t_B). Spatiotemporal characteristics of upper-limb reaching movements during VR test performance were reconstructed from motion capture data. Statistics included correlations and repeated measures analysis of variance.

Results: Construct validity was substantiated by moderate correlations between the'gold standard' pencil-and-paper CTT and the VR adaptations (DOME-CTT: t_A 0.58, t_B 0.71; HMD-CTT: t_A 0.62, t_B 0.69). VR versions showed relatively high test–retest reliability (intraclass correlation; VR: t_A 0.60–0.75, t_B 0.59–0.89; original: t_A 0.75–0.85, t_B 0.77–0.80) and discriminant validity (area under the curve; VR: t_A 0.70–0.92, t_B 0.71–0.92; original: t_A 0.73–0.95, t_B 0.77–0.95). VR completion times were longer than for the original pencil-and-paper test; completion times were longer with advanced age. Compared with Trails A, Trails B target-to-target VR hand trajectories were characterized by delayed, more erratic acceleration and deceleration, consistent with the greater executive function demands of divided vs. sustained attention; acceleration onset later for older participants.

Conclusions: The present study demonstrates the feasibility and validity of converting a neuropsychological test from two-dimensional pencil-and-paper to three-dimensional VR-based format while preserving core neuropsychological task features. Findings on the spatiotemporal morphology of motor planning/execution during the cognitive

*Correspondence: Meir.PlotnikPeleg@sheba.health.gov.il
†Oran Ben-Gal and Glen M. Doniger contributed equally as co-second authors on this manuscript
[1] Center of Advanced Technologies in Rehabilitation, Sheba Medical Center, Ramat Gan, Israel
Full list of author information is available at the end of the article

tasks may lead to multimodal analysis methods that enrich the ecological validity of VR-based neuropsychological testing, representing a novel paradigm for studying cognitive-motor interactions.

Keywords: Executive functions, Cognitive-motor interactions, Construct validity, Divided attention, Neuropsychological testing, Virtual reality

Background

The term "executive functions" is an umbrella term for a wide range of cognitive processes and behavioral competencies necessary for the cognitive control of behavior including problem solving, planning, sequencing, sustained attention, utilization of feedback, and multitasking [1]. Neuropsychological tests of executive functions aim to assess these processes [2]. Accordingly, performance on these tests is assumed indicative of executive functioning in everyday living [3]. One of the limitations of these tests relates to their low 'ecological validity', namely the uncertainty about how closely they reflect capacity of executive function in real life [4–6]. In this regard, Burgess et al. [7] has claimed that "the majority of neuropsychological assessments currently in use were developed to assess 'cognitive constructs' without regard for their ability to predict 'functional behavior'."

Neuropsychological assessment in virtual reality (VR)

Early discussions of ecological validity in neuropsychology emphasized that the technologies available at that time could not replicate the setting in which the behavior of interest actually occurs [8]. Furthermore, currently, most neuropsychological assessments still use outdated methods (e.g., pencil-and-paper administration; static stimuli) that have yet to be validated with respect to real-world functioning [9].

To overcome this limitation, testing participants in real word situations (e.g., the Multiple Errands Test [MET] [10]) has been considered an ecologically valid and advantageous alternative to traditional tests [11]. However, this approach is logistically challenging, requiring travel to a naturalistic testing site [12].

In an attempt to overcome this logistical hurdle, the Virtual Errands Test (VET) was devised by McGeorge et al. [13] as an adaptation of the MET for VR-based administration. Still, this test, and similar VR variants, are limited in their ability to distinguish between healthy and clinical cohorts (see [11] for a review) and to yield performance on the virtual tasks similar to performance in the real world (e.g., [14, 15]). Further, most VR-based tests like VET involve presenting a simulated VR environment on a standard computer screen (e.g., Elkind et al. [16]), which may lead to a non-immersive experience, thus paradoxically compromising rather than enhancing ecological validity.

Notably, VR-based tests simulating shopping tasks for the assessment of executive function have demonstrated good ecological validity [17, 18]. However, the approach of adapting executive function testing for the VR environment has not been widely accepted in both research and clinical contexts.

Research rationale

Critically, we posit that the concept of 'ecological validity' is not merely related to the type of task performed and its relevance to daily living. In general, each response on a cognitive task involves interactions with sensory and motor functions, first to determine the required behavioral response and then to plan and execute it. These processes cannot be distinguished and examined with traditional pencil-and-paper testing or even with computerized testing platforms.

Thus, as a first step, we aim to develop VR neuropsychological tests by adapting well-validated traditional neuropsychological tests that measure particular cognitive constructs. These adaptations will enhance ecological validity by including multi-multimodal (e.g., cognitive-sensory-motor) interactions, facilitating measurement of cognitive function in a manner more relevant to to the interaction among multiple functions characteristic of everyday activities [19–24]. Specifically, the VR technology we employ allows for collection of quantitative three-dimensional kinematic data (unavailable for traditional neuropsychological tests) that tracks motion in space and may improve our ability to define and discriminate among levels of performance.

The Color Trails Test (CTT)

The Trail Making Test (TMT) [25, 26] is among the most popular pencil-and-paper tests of executive function, attention and processing speed in research and clinical neuropsychological assessment. The Color Trails Test (CTT) is a culture-fair variant of the TMT. In *Trails A* the participant draws lines to sequentially connect circles numbered 1–25 (odd-numbered circles are pink; even-numbered circles are yellow). In *Trials B* the participant alternates between circles of two different colors (i.e., 1-pink, 2-yellow, 3-pink, 4-yellow, etc.) [27]. Scoring is based on the time needed to complete the tasks, with shorter time reflecting better performance. It has been proposed that *Trails A* assesses sustained visual attention involving perceptual tracking and simple sequencing, while Trails B more directly assesses executive function processes, including divided attention, simultaneous alternating and sequencing [27, 28].

The present study

The overall goal of this study was to demonstrate the value of adapting a well-validated paper-and-pencil executive function task for VR administration. We developed two VR adaptations of the CTT test: (i) the DOME-CTT, designed for a large-scale VR system, in which the stimuli are projected on a 360° dome-shaped screen surrounding the participant, and (ii) the HMD-CTT, designed for a low-cost head-mount device (HMD), in which the stimuli are presented via VR goggles. In addition to developing the VR-based tests, we evaluated their ability to measure the same cognitive constructs (construct validity) as the gold standard pencil-and-paper CTT, as well as their ability to differentiate among healthy young, middle-aged and older age groups (discriminant validity) relative to the original CTT. Finally, we explored cognitive-motor interactions during performance of the VR-CTT tasks.

Methods

General

Two VR-CTT platforms were developed: DOME-CTT and HMD-CTT. Findings from experiments using these platforms are described in Study 1 and Study 2, respectively. There were a total of 147 healthy participants in Study 1 and Study 2 who completed this testing as part of larger experimental protocols (see Additional file 1: Table S1). Participants were subdivided into the following age groups: (1) young adults (YA), ages 18–39 years (n = 50); (2) middle-aged adults (MA) ages 40–64 years (n = 80); and (3) older adults (OLD), ages 65–90 years (n = 17). For all groups, exclusion criteria were motor, balance, psychiatric or cognitive conditions that may interfere with understanding the instructions or completing the required tasks (determined by screening interviews). The protocols were approved by the Sheba Medical Center institutional review board (IRB), and all participants signed informed consent prior to enrolling in the study.

Methods for Study 1 (DOME-CTT)

Participants

Data from 14 YA [age: 27.9 ± 5.0 (mean ± SD) years, education: 16.4 ± 2.9 (mean ± SD) years; 9 females] and 29 MA (age: 55.8 ± 6.2 years, education: 16.3 ± 3.0 years; 16 females) were included in Study 1.

Apparatus

A fully immersive virtual reality system (CAREN High End, Motek Medical, The Netherlands) projected a virtual environment consisting of the task stimuli on a full-room dome-shaped screen surrounding the participant (Fig. 1). The system comprises a platform with an embedded treadmill and is synchronized to a motion capture system (Vicon, Oxford, UK). Auditory stimuli and feedback are delivered via a surround sound system.

Adapting the pencil-and-paper Color Trails Test for large-scale VR—The DOME-CTT (Fig. 2)

A virtual version of the CTT was developed to demonstrate the feasibility of performing neuropsychological testing in a virtual environment. The original pencil-and-paper CTT consists of four parts: practice (Trails) A, test (Trails) A, practice (Trails) B and test (Trails) B [27]. As below, all were adapted to the VR environment. In the VR version of the CTT, the two-dimensional (2D) page (Fig. 2a) is replaced with a three-dimensional (3D) VR space (Fig. 2b, c) that introduces the dimension of depth to the target balls (that replace the 2D circles) and to the generated trajectory. The translation to 3D geometry followed principles governing the 2D design (compare Fig. 2a and Fig. 2b, c). For example: (1) balls were positioned so that virtual trajectories between sequential

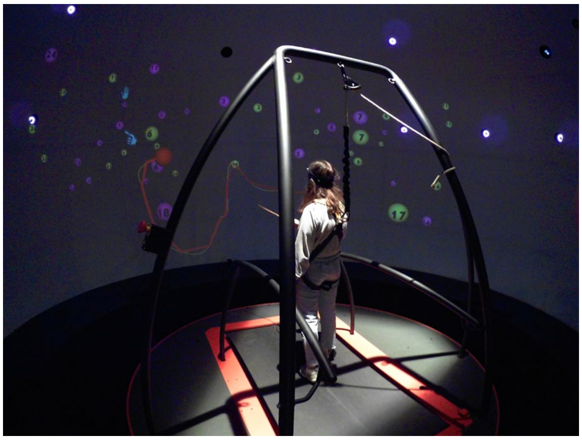

Fig. 1 The CAREN High End (Motek Medical, The Netherlands) has a 6-degree of freedom (three translation axes, three rotation axes) moveable platform that is synchronized with a virtual visual scene projected on a 360° dome-shaped screen (projection resolution 1920X1080 lines). A split-belt treadmill is embedded in the platform (not operated in this study). A motion capture system (Vicon) and a pair of force plates embedded in the treadmill provide data on the participant's spatiotemporal position during the task (sampling rate 120 Hz). The participant who performs the DOME-CTT VR adaptation (see text), is secured with safety harness, holding a wand-like pointing stick with a marker on its edge in order to control the avataric red ball and to move it towards the target ball

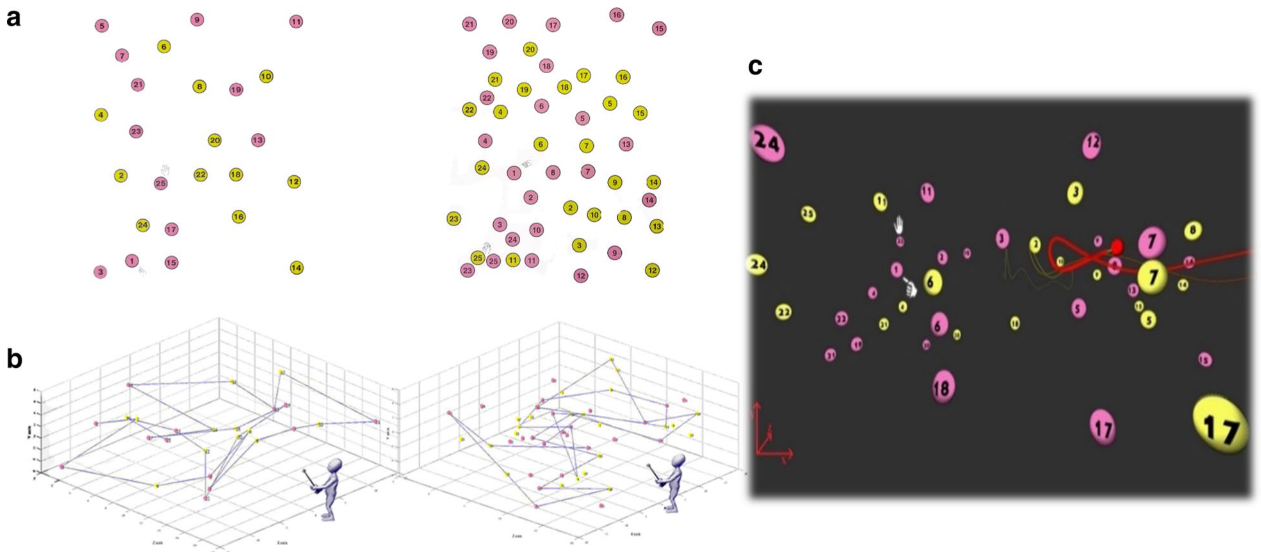

Fig. 2 Adaptation of the pencil-and-paper CTT to yield the DOME-CTT. **a** Trails A (left) and B (right) tasks from the pencil-and-paper CTT. The first and last balls are each indicated by a picture of a hand (pointing and stop signing, respectively, seen here in Trails B only). **b** Schematic depiction of the spatial geometric orgnizaiton of the target balls in the three dimensional space. **c** Trails B task from the DOME-CTT, as viewed by the participant. Ball size varies as a function of its proximity to the participant in the VR space. A fixed pointing stick marker is represented by the red ball. Its trace is depicted in yellow, while the red, more pronounced trace is used for visual feedback in the VR space. The first and last balls are indicated using the hand icon

target balls would not cross previous trajectories (i.e., between target balls from earlier in the task sequence); (2) proximity of balls in a given region of the 3D space was similar to that in the corresponding region of 2D space in the original CTT; (3) for Trails B, we positioned the corresponding identically-numbered distracter ball of incorrect color at a relative distance to the target ball similar to the that in the original 2D CTT.

The participant performed the DOME-CTT with a marker affixed to the tip of a wand-like pointing stick held in the dominant hand (corresponding to the pen or pencil in the original CTT). The three-dimensional coordinates of the marker were tracked in real time by the motion capture system at a sampling rate of 120 Hz. A virtual representation of this marker appeared within the visual scene (i.e., 'avatar', represented by a small red ball—Fig. 2c). To mimic drawing lines in the 2D pencil-and-paper CTT, as the participant moved his/her hand within the VR space, a thick red 'tail' trailed directly behind the position of the (red ball) avatar, gradually becoming a faint yellow tail as the avatar moved farther away from the initial position (Fig. 2c).

Movement of the marker was recorded in real time by a motion capture system that allows the reconstruction of kinematic data over the duration of the test.

The testing procedure was also adapted for the new format. As above, the original pencil-and-paper CTT comprises four consecutively administered test levels: (1) Trails A practice; (2) Trails A; (3) Trails B practice; and (4) Trails B [16]. Though drawing lines with a pen/pencil on a piece of paper is highly familiar, manipulation of the VR 'controller' (i.e., the marker affixed to the pointing stick) to move an avatar (i.e., the red ball) within the virtual environment is a relatively unfamiliar skill. Thus, the DOME-CTT began with an additional practice level in which participants practiced guided movement of the avatar within the virtual space to so that it touched the numbered ball targets. During this level, participants were introduced to the positive feedback received when the avatar ball touched the correct ball (i.e., momentary enlargement of the ball) and the negative feedback when it touched an incorrect ball (i.e., brief buzzing sound). These feedback stimuli were also presented during the remainder of the testing session. After this initial practice level, test levels corresponding to those in the original CTT were administered. However, unlike the pencil-and-paper CTT, Trails A and Trails B were each preceded by two different practice levels. In the first practice level, all virtual balls were clustered near the center of the visual field, and in the second practice level, the balls were distributed throughout the visual field, approximating the spatial distribution of the balls in the actual testing levels. A video demonstration of the DOME-CTT is provided in Additional file 2.

Procedure

Data on pencil-and-paper CTT and DOME-CTT were collected as part of three different experimental protocols (see Additional file 1: Table S1). All data (with the exception of test retest data) described in this study were collected on the first visit. The participants completed the pencil-and-paper CTT and DOME-CTT on the same day in counterbalanced order across participants. We monitored the general wellbeing of the participants (e.g., absence of fatigue) throughout the tests.

Outcome measures and statistical analysis

For the pencil-and-paper CTT and the DOME-CTT, completion times for Trails A and B were recorded (t_A, $t_{B,}$ respectively). Construct validity was assessed by correlating t_A and t_B from the DOME-CTT with the corresponding scores from the gold standard CTT (Pearson coefficient). Analysis of variance (ANOVA) was used to assess effects of Group (young, middle aged; between-subjects factor), Trails (Trails A, Trails B; within -subjects factor) and Format (pencil-and-paper CTT, DOME-CTT; within-subjects factor). Partial Eta Squared was computed as a measure of effect size. To verify suitability of parametric statistics, Shapiro–Wilk normality tests were run for each outcome variable per group. Of the eight normality tests, none indicated non-normal distributions (Shapiro–Wilk statistic ≤ 0.93; $p \geq 0.16$). Levene's test [29] revealed inhomogeneity of variance among groups for Trails A and B in pencil-and-paper and VR formats ($p < 0.05$). Therefore, the data were log-transformed prior to applying ANOVA tests. On the new data sets we confirmed homogeneity of variance assumption for Trails A and B of the pencil-and-paper CTT and for Trails B of the DOME-CTT ($p > 0.05$). Descriptive statistics, figures and correlations analyses were performed on the pre transformed data.

Summary statistics (mean ± SD) were computed for t_A and t_B from the pencil-and-paper CTT and DOME-CTT.

Errors were manually recorded by the experimenter for the pencil-and-paper CTT [27]; for the DOME-CTT, errors were recorded both manually and automatically by the software. Related samples Wilcoxon Sign Test (non-parametric) test was used to evaluate the Format effect separately for Trails A and B. Mann–Whitney U tests were used to evaluate the group effect.

To examine discriminant validity (i.e., ability to separate between YA and MA) of the DOME-CTT as compared with the pencil-and-paper CTT, we plotted receiver operating characteristic curves (ROC) for Trails A and Trails B (i.e., t_A and $t_{B,}$ respectively) for each test format and calculated the area under the curve (AUC; range: 0–1, higher values reflect better discriminability).

Level of statistical significance was set at 0.05. Statistical analyses were run using SPSS software (SPSS Ver. 24, IBM).

Methods for Study 2

Participants

Data from 36 YA (age: 26.7 ± 4.1 [mean ± SD] years, education: 15.9 ± 2.3 [mean ± SD]; 21 females), 51 MA (age: 56.2 ± 6.2 years, education: 16.8 ± 3.0 years; 39 females) and 17 OLD (age: 73.7 ± 6.5 years, education: 13.1 ± 2.7 years*;* 11 females) were included in Study 2.

Apparatus

VR technologies have advanced rapidly in recent years. In addition to new technical features for precise stimulus delivery and response measurement, as well as enhanced usability, low-cost VR is now widely accessible. The most accessible type of VR system is the head-mount device (e.g., HTC Vive, Oculus Rift), which is designed for home-based operation. In addition to its continued popularity for entertainment, VR is now being applied in a variety of 'serious' contexts, ranging from surgical simulation to the study of human performance and psychological function [19, 30]. For this study, we used a fully immersive VR system (HTC-Vive; New Taipei City, Taiwan) including a headset with ~ 100° field of view (FOV) in the horizontal plane and ~ 110° FOV in the vertical plan. Also included were a controller for user interaction with the virtual environment and two 'lighthouse' motion trackers for synchronizing between actual controller position and corresponding position in the virtual environment.

Adapting the pencil-and-paper Color Trails Test for a headset-based VR system—The HMD-CTT

In developing the HMD-CTT version, we adopted a similar approach to the development of the DOME-CTT. Briefly, we used the popular Unity3D VR game engine [31] to design a virtual environment for the CTT. With the exception that the participant held the HTC controller rather than a wand-like pointing stick, task design matched the DOME-CTT, including positive and negative feedback cues, practice and test procedures. Figure 3 illustrates how the 2D format of the original CTT was translated to the 3D HMD-CTT format (Trails A; see also video demonstration in Additional file 3). The HMD-CTT incorporated practice and test levels corresponding to those in the DOME-CTT described above.

Procedure

The procedure was identical to that of Study 1 (see above).

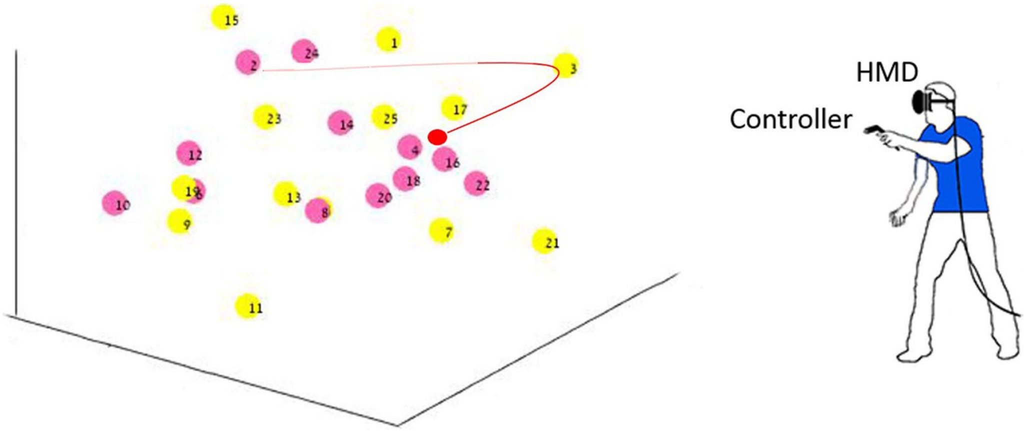

Fig. 3 Translation of pencil-and-paper CTT to HMD-based VR format. In the three-dimensional VR-based HMD-CTT (HTC-Vive, projection resolution 2160X1200 lines), the participant completes the task while standing and wearing a VR headset. S/he holds a VR controller in his/her dominant hand to move the (red ball) avatar to sequentially connect numbered balls distributed throughout the virtual space visible in the headset. (Trails A shown.)

Outcome measures and statistical analyses

For the pencil-and-paper CTT as well as the HMD-CTT, completion times for Trails A and B were recorded (t_A, $t_{B,}$ respectively). Similar to Study 1, construct validity of t_A and t_B was assessed by correlating t_A and t_B from the HMD-CTT with the corresponding scores from the gold standard pencil-and-paper CTT. As in the DOME-CTT study, repeated-measures ANOVA was used to assess the effects of Group (YA, MA, OLD; between-subjects factor), Trails (Trails A, Trails B; within-subjects factor) and Format (pencil-and-paper CTT, HMD-CTT; within-subjects factor). Partial Eta Squared was computed as a measure of effect size. We used the Bonferroni correction to adjust for multiple comparisons in the post-hoc pairwise comparisons.

As for Study 1, Shapiro–Wilk normality tests were run for each outcome variable per group to verify the suitability of parametric statistics. Of the twelve normality tests, four indicated non-normal distributions: t_A pencil-and-paper CTT, MA group, Shapiro–Wilk statistic = 0.786, $p = 0.002$; t_B pencil-and-paper CTT, MA group, Shapiro–Wilk statistic = 0.770, $p = 0.002$; t_B pencil-and-paper CTT, YA group, Shapiro–Wilk statistic = 0.744, $p = 0.001$, t_A pencil-and-paper CTT, OLD group, Shapiro–Wilk statistic = 0.844, $p = 0.015$. As in Study 1, Levene's test [29]. revealed inhomogeneity of variance among groups for Trails A and B in pencil-and-paper and VR formats ($p < 0.05$). Thus, the data were log-transformed prior to applying ANOVA analyses, and homogeneity of variance was confirmed for paper-and-pencil Trails A and B ($p > 0.05$).Descriptive statistics, figures and correlations analyses were performed on the pre transformed data. Additional analyses are described in the *Results* section.

Regarding prevalence of errors, related samples Wilcoxon Sign Test (non-parametric) test was used to evaluate the Format effect separately for Trails A and B. Kruskal–Wallis tests were used to evaluate the group effect (three levels) separately for Trails and Format.

As for Study 1, level of statistical significance was 0.05, and statistic analyses were run with SPSS.

Qualitative analysis of manual performance

Spatial coordinates of the controller position (corresponding to 'red ball' avatar) were recorded throughout the HMD-CTT sessions. Custom software written in MATLAB (Mathworks, Inc.) used this data to extract and analyze the 24 target-to-target reaching movements during Trails A and Trails B, respectively (errors were excluded from this analysis). We made a qualitative assessment of the trajectories generated in each of the three groups by examining the grand averages of their velocity profiles to characterize upper-limb motor behavior associated with the HMD-CTT tasks. For a full description of the methodology used to generate these grand averages, see Additional file 1.

Evaluation of test–retest reliability (Study 1 and Study 2)

To evaluate test–retest reliability, some participants completed a second assessment.

Fifteen MA participants from Study 1 completed an additional evaluation about 12 weeks after the initial evaluation (per protocol 1 in Additional file 1: Table S1) during which they completed the pencil-and-paper

CTT and DOME-CTT in the same order as in the initial evaluation.

Thirty-two MA participants from Study 2 completed an additional evaluation about 12 weeks after the initial evaluation. Also from Study 2, twenty participants ($n=10$ YA, $n=1$ MA, $n=9$ OLD) completed an additional evaluation 2 weeks after the initial one. The pencil-and-paper CTT and HMD-CTT were administered in the same order as in the initial evaluation.

To assess test–retest reliability, we computed intraclass correlation coefficients (ICC; two-way mixed, effects, absolute agreement, [32]) for t_A and t_B scores from the traditional pencil-and-paper CTT and the DOME-CTT (*Study 1*) or HMD-CTT (Study 2) collected at two visits. ICC reflects similarity of the obtained scores irrespective of the level of performance, reflecting not only correlation, but also agreement between measurements [33, 34]. By convention ICC > 0.75 is considered good reliability [32].

Results

Study 1

Performance on the DOME-CTT: group, trails and format effects

All participants completed the tests in both formats. Time for initial DOME-CTT practice levels varied between participants, but usually did not exceeded 10–15 min. Due to technical malfunction, data for DOME-CTT Trails A from one YA participant was not recorded.

Statistical analysis revealed effects of Group ($F_{1,40}=25.4$, $p<0.001$, $\eta^2=0.38$; longer completion time for middle-aged), large effects of Trails ($F_{1,40}=273.7$, $p<0.001$, $\eta^2=0.87$; longer completion time for Trails B) and large effects of Format (t_A: $F_{1,40}=1301.7$, $p<0.001$, $\eta^2=0.97$; longer completion time for DOME-CTT). The Format X Trails interaction was also found to be significant ($F_{1,40}=19.5$, $p<0.001$, $\eta^2=0.32$,), reflecting a larger difference between Trails A and B for the DOME-CTT (i.e., as compared to the pencil-and-paper CTT. None of the other interactions (i.e., Group X Format, Group X Trails X Format and Group X Trails) were statistically significant ($p\geq0.09$).

Errors were more prevalent during performance of the DOME-CTT (Fig. 4) for both Trails A and B ($p<0.001$). The group effect did not reach statistical significance ($p\geq0.056$). The MA group made more errors than the YA group irrespective of Trails and Format ($p\leq0.002$).

Correlations between pencil-and-paper and DOME-CTT completion times

Figure 5 shows the relationship between performance on the pencil-and-paper and DOME-CTT for the YA (blue) and MA (orange) groups. The Spearman correlation (rho; r_s) between Part A completion time (t_A) on the gold-standard pencil-and-paper CTT and the corresponding Part A completion time on the DOME-CTT was 0.58 ($p<0.001$; Fig. 5a). For Part B completion time (t_B), the Spearman correlation was 0.70 ($p<0.001$; Fig. 5b).

Study 2

Performance on the HMD-CTT: group, trails and format effects

One OLD participant (age: 70 years, education: 12 years, female) could not perform the HMD-CTT practice and actual test levels, showing general disorientation. Another participant from this group (age: 89 years, education: 12 years, male) asked to stop the HMD-CTT during the Trails B portion, expressing frustration at his perceived poor performance (see also Fig. 7 legend). Data from these participants were omitted from relevant analyses.

Statistical analysis revealed significant effects of Group ($F_{2,98}=41.9$, $p<0.001$, $\eta^2=0.46$; progressively longer completion time with more advanced age, all pairwise comparisons $p<0.001$), Trails ($F_{1,98}=617.2$, $p<0.001$, $\eta^2=0.86$; longer completion time for Trails B) and Format ($F_{1,98}=660.8$, $p<0.001$, $\eta^2=0.87$; longer completion time for HMD-CTT). Of the interaction effects, only the Format X Trails interaction was significant ($F_{2,98}=14.3$, $p<0.001$, $\eta^2=0.12$) due to a larger Trails effect for the HMD-CTT vs. the pencil-and-paper CTT. The Group x Format, Group X Trails and Group X Format X Trails interactions were not significant ($p>0.30$).

As participants in the OLD group had significantly fewer years of education than participants in the YA and MA groups ($p\leq0.001$; see *Methods*), we repeated the analysis entering years of education as a covariate. The results did not change appreciably (see Additional file 1).

Errors were more prevalent during performance of the HMD-CTT (Fig. 6, $p\leq0.004$). Interestingly, a significant group effect was found only for HMD-CTT (Trails A and B; $H\geq8.8$, $p\leq0.012$) but not for the pencil-and-paper CTT ($H\leq2.96$; $p\geq0.227$). Post-hoc analyses revealed that the significant Group effect for HMD-CTT was attributable to more errors among the OLD than the YA ($p=0.009$).

Correlations between pencil-and-paper and HMD-CTT—completion times

Figure 7 shows the relationship between performance on the pencil-and-paper and HMD-CTT for the YA (blue), MA (orange) and OLD (green) group. The Spearman correlation (rho; r_s) between Part A completion time (t_A) on the gold-standard pencil-and-paper CTT and the

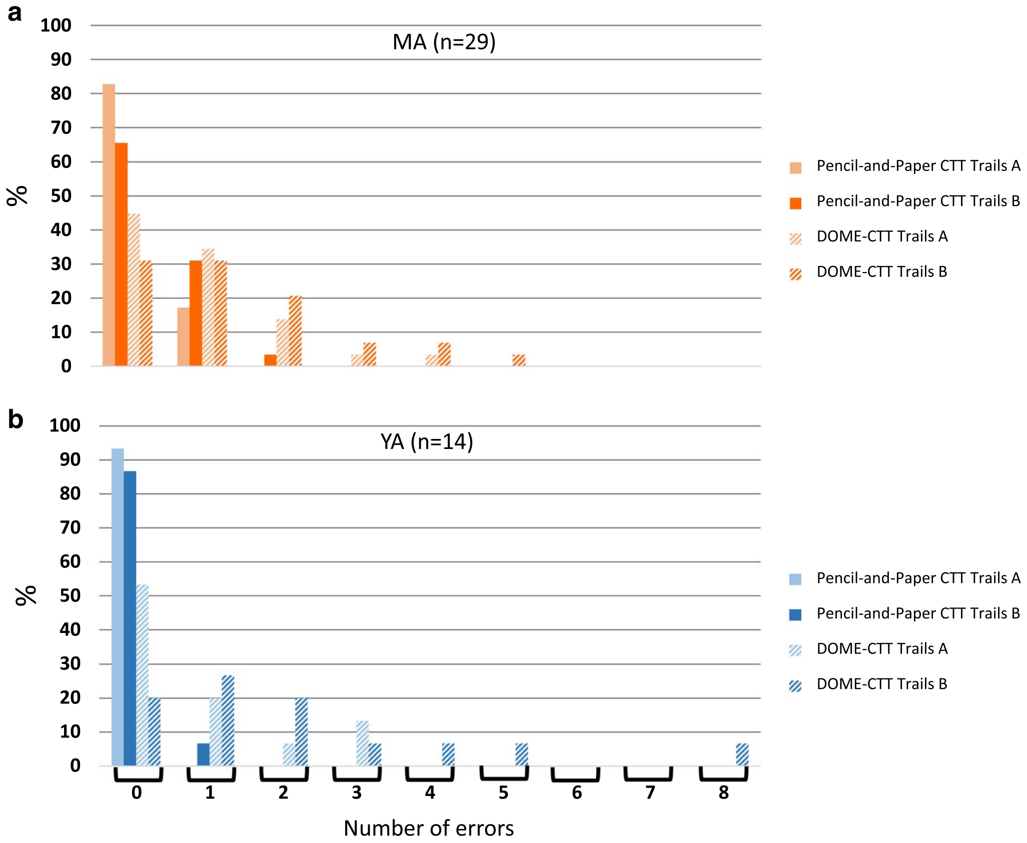

Fig. 4 Comparison of error rates between pencil-and-paper CTT and DOME-CTT VR-based adaptation. Percent of participants making 1, 2, 3, 4, 5, 6, 7, or 8 errors for paper-and-pencil and DOME-CTT versions of Trails A and Trails B tasks (see key), separately for the MA group (**a**) and the YA group (**b**). For the pencil-and-paper CTT, most participants in both groups made no errors, and none made more than 2 errors. Conversely, for the DOME-CTT, less than half of the participants (YA and MA combined) made no errors and some made a substantial number of errors

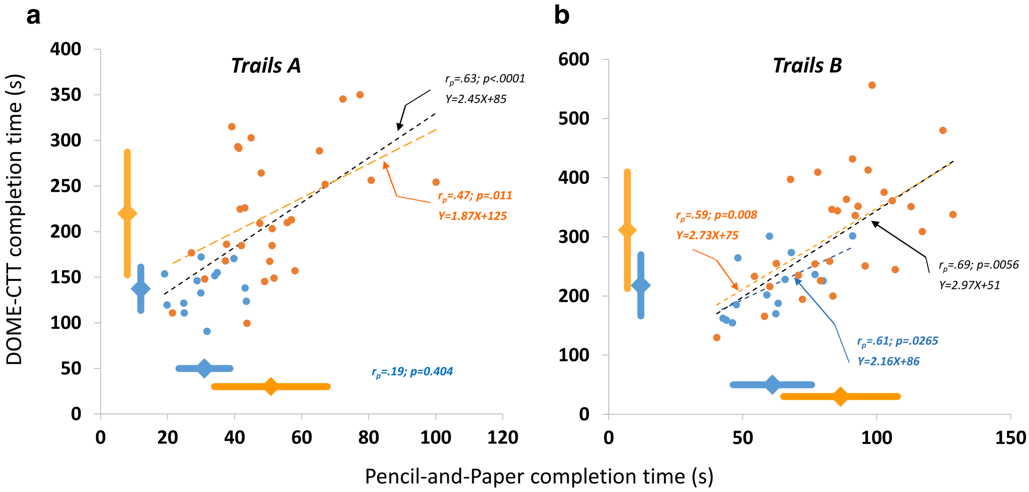

Fig. 5 Convergent construct validity of the VR-based DOME-CTT. Trails A (t_A, left panel) and Trails B (t_B, right panel) completion time recorded during the gold-standard pencil-and-paper CTT plotted against the corresponding completion times recorded during the DOME-CTT for YA (blue) and MA (orange) participants. One YA datapoint is missing for Trails A, as the participant did not complete the DOME-CTT (i.e., due to technical malfunciton). Pearson correlations are shown for YA, MA, as well as combined (black) groups, and regression lines are plotted for significant correlations. Diamond markers and thick lines adjacent to the axes indicate mean ± SD for YA and MA groups

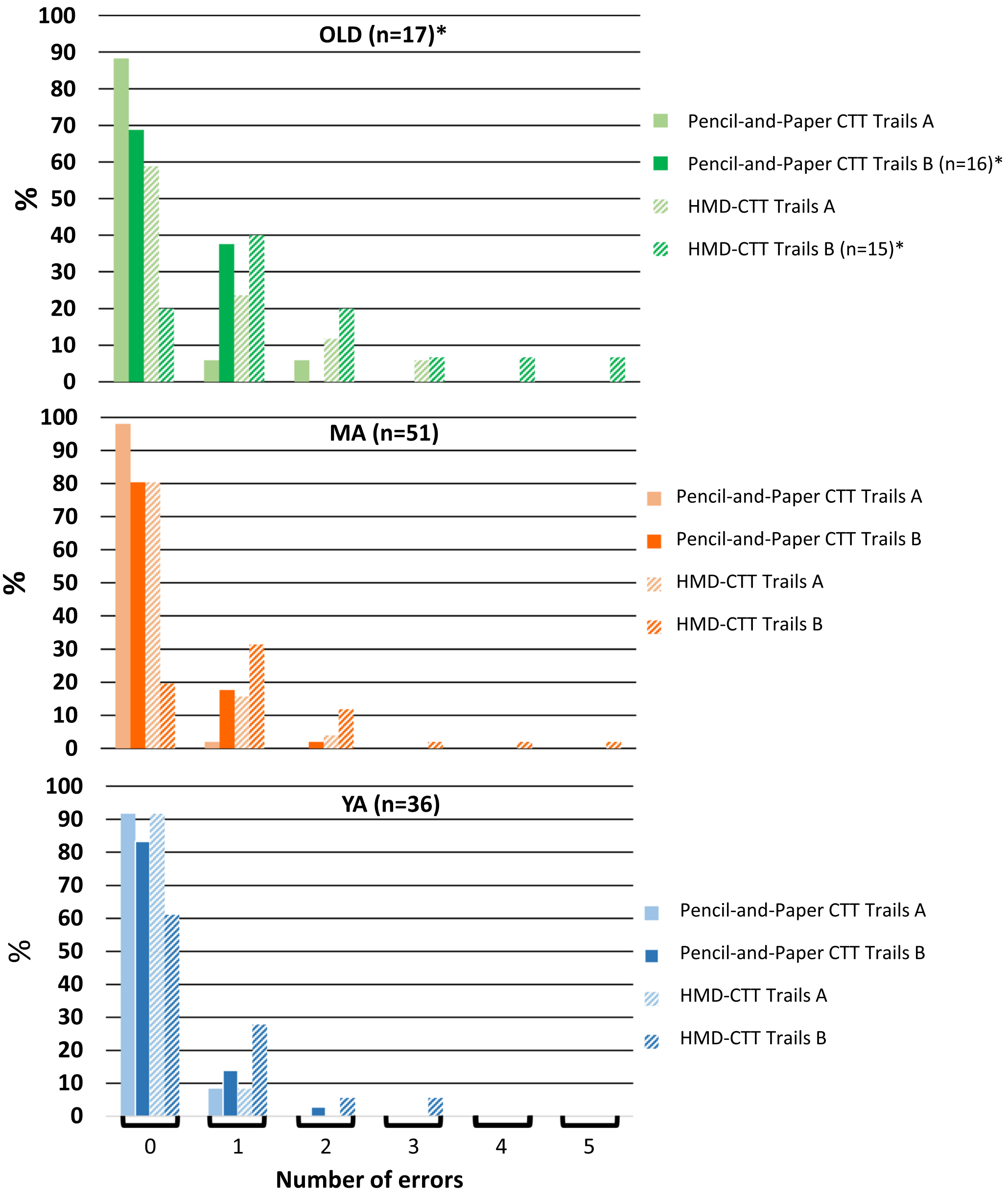

Fig. 6 Comparison of error rates between pencil-and-paper CTT and HMD- CTT VR-based adaptation. Percent of participnats making 1, 2, 3, 4, or 5 errors for paper-and-pencil and HMD-CTT versions of Trails A and Trails B tasks (see key), separately for OLD (top panel), MA (middle panel), and YA (bottom panel) groups. Similar to the DOME-CTT (Fig. 5), substantially more errors were made for the VR-based HMD-CTT than for the paper-and-pencil CTT; this tendency was particularly evident for OLD participants

corresponding Part A completion time on the HMD-CTT was 0.62 ($p < 0.001$; Fig. 7a). For Part B completion time (t_B), the Spearman correlation was 0.69 ($p < 0.001$; Fig. 7b).

Qualitative analysis of manual performance

Figure 8 shows grand averages of the scaled ball-to-ball hand trajectory velocity profiles (see methodologies in Additional file 1) for all participants who completed HMD-CTT Trails A (solid traces) and B (dashed traces). Data for the three age groups is color-coded (see legend for details). Based on these traces, the following observations can be made:

(1) For all groups and for both Trails A and B, movement toward a target (ball) does not stop immediately upon reaching the target (virtually touching, at x = 100% which for all, except for the last trajec-

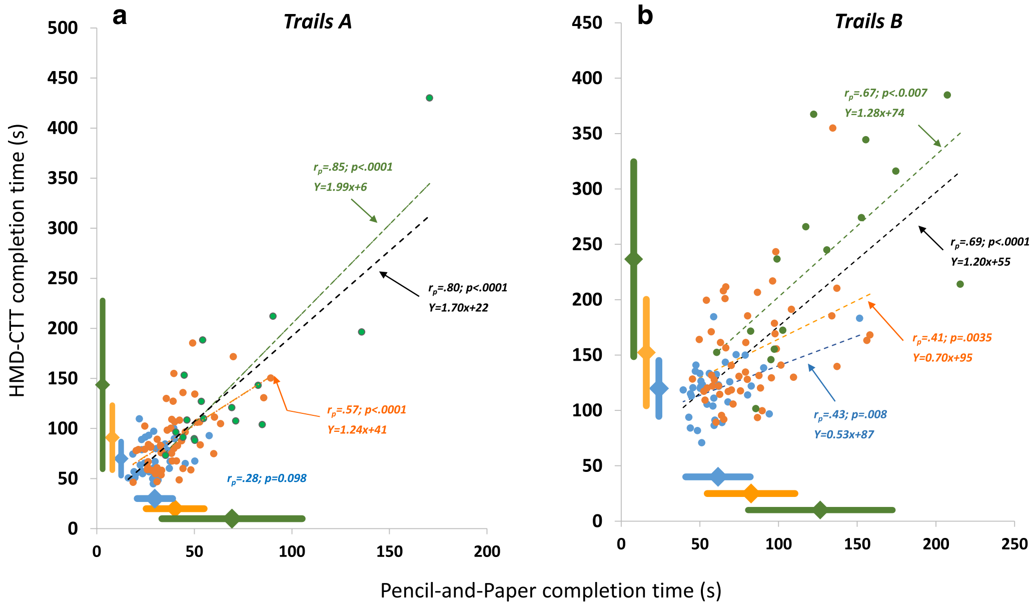

Fig. 7 Convergent construct validity of the VR-based HMD-CTT. Trails A (t_A, left panel) and Trails B (t_B, right panel) completion time recorded during the gold-standard pencil-and-paper CTT plotted against the corresponding completion times recorded during HMD-CTT, for YA (blue), MA (orange), and OLD (green) participants. Two OLD datapoints are missing as the participants did not complete the HMD-CTT.*Pearson correlations are shown for YA, MA, OLD and combined (black) groups, and regression lines are plotted for significant correlations. Dimond markers and thick lines adjacent to the axes indicate mean ± SD for YA, MA, and OLD groups. *One OLD participant who could not finish the HMD-CTT Trails B had the slowest Trails A completion time, for both pencil-and-paper (i.e., 170.5 s) and HMD versions (450.2 s). His completion time for the pencil-and-paper Trails B was 327.3 s (this data point was not plotted)

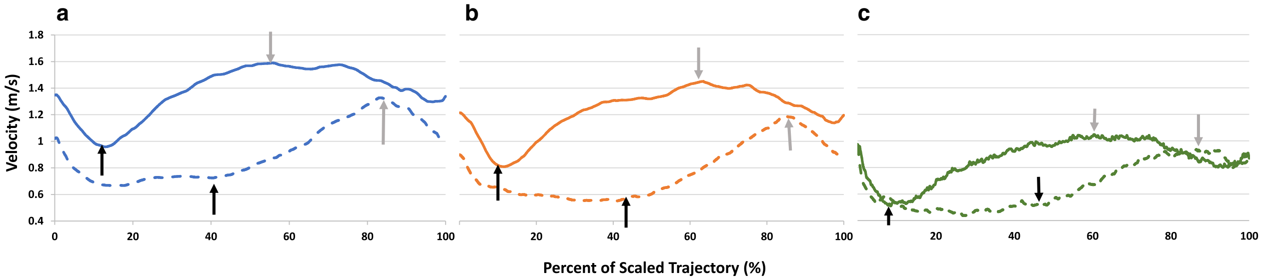

Fig. 8 Grand averages of HMD-CTT hand movement velocity profiles. HMD-CTT hand movement velocity (meters/second) profiles showing trajectories for young (**a**), middle-aged (**b**), and older adult (**c**) groups for Trails A (solid line) and Trails B (dashed lines) over time, scaled to reflect percent from trajectory completion duration. In **a** (YA), for each of Trails A and B, 840 trajectories were avaraged from 35 participants; in **b** (MA), 1176 trajectories were averaged from 49 participans; in **c** (OLD), 336 trajectories were averaged from 14 participants. See *text* for additional description and interpretation

tory, is also x = 0% of the next trajectory), but little later, as reflected by the ensuing gradual decrease in velocity decrease at x = 0%, reaching a minimum and ensuing gradual increase in velocity. This initial decrease on the grand average traces does not reach zero because the minimum is reached at a different time for each of the 24 individual ball-to-ball trajectories (see Additional file 1: Fig. S1).

(2) For Trails A, but not for Trails B, soon after reaching this minimum (i.e., completing the execution of the previous trajectory), a new trajectory can be identified (time of emergence is designated by the left black arrow). The velocity profile of this trajectory is characterized by an accelerating portion (peak indicated by a gray arrow) and a decelerating portion upon approaching the target ball. The

Table 1 Mean values for CTT completion times (mean ± SD)

	Pencil-and-Paper CTT	DOME-CTT
YA: t_A (s)	31.01 ± 7.62	137.39 ± 23.84
YA: t_B (s)	61.09 ± 14.66	213.16 ± 53.18
MA: t_A (s)	50.81 ± 16.80	219.87 ± 67.31
MA: t_B (s)	86.53 ± 21.22	311.26 ± 98.61

YA Young adults; *MA* Middle Aged, *s* seconds; t_A completion time of Trails A; t_B completion time of Trails B

Table 2 Mean values for CTT completion times (mean ± SD)

	Pencil-and-Paper CTT	HMD-CTT
YA: t_A (s)	29.78 ± 9.08	69.93 ± 17.03
YA: t_B (s)	61.52 ± 20.53	119.82 ± 25.23
MA: t_A (s)	40.15 ± 15.02	91.36 ± 32.60
MA: t_B (s)	82.60 ± 27.88	152.24 ± 47.99
OLD: t_A (s)	63.55 ± 35.95	126.75 ± 84.26
OLD: t_B (s)	126.57 ± 45.92	236.52 ± 88.00

YA Young adults, *MA* Middle Aged, *EA* Elderly adults, *s* seconds; t_A completion time of Trails A; t_B completion time of Trails B

degree of asymmetry between these portions of the trajectory varies between groups, with YA showing greater symmetry.

(3) Conversely, for Trails B, a prolonged 'executive function' period is evident, and acceleration toward the target is identifiable after at least 40% of the trace, with the older groups showing a more delayed initiation of movement (black arrows on dashed traces). This pattern is consistent with the divided attention aspect of Trails B, in which the participant must locate the ball next in the numerical sequence but of opposite color to the current ball, ignoring the distracter ball with the correct number but the incorrect color (i.e., same color as the current ball).

(4) Consistent with the results for completion times (t_A, t_B; Table 2), the velocity profiles for Trails A are faster than those of Trails B.

Test–retest reliability (Studies 1 and 2)

For the DOME-CTT (retest interval of about 12 weeks), moderate reliability was found: Trails A (t_A) and B (t_B) (ICC = 0.597, $p = 0.023$, ICC = 0.676, $p = 0.018$, respectively). For the pencil-and-paper CTT, comparatively better reliability was found for Trails A (t_A) (ICC = 0.778, $p = 0.001$) and Trails B (t_B) (ICC = 0.622, $p = 0.040$).

For the HMD-CTT (retest interval of about 12 weeks), moderate reliability was found for Trails A (t_A) and B (t_B) (ICC = 0.618, $p = 0.004$; ICC = 0.593, $p = 0.003$, respectively). For the pencil-and-paper CTT, comparatively better reliability was found for Trails A (t_A) and Trails B (t_B) (ICC = 0.744, $p < 0.001$; ICC = 0.769, $p < 0.001$, respectively).

In both VR-CTT formats, there were no substantial differences in ICC values for MA participants who engaged in cognitive training during the 12 week period (as part of the larger study protocol; see Additional file 1: Table S1) as compared with those who did not (see details in section D of Additional file 1).

For the HMD-CTT (retest interval of about 2 weeks), good reliability was found for Trails A (t_A) and B (t_B) (ICC = 0.748, $p = 0.002$; ICC = 0.893, p < 0.001, respectively). The paper-and-pencil CTT also showed good reliability for both Trails A (t_A) and Trails B (t_B); Compared to HMD-CTT, the gold standard test had higher reliability for t_A and lower reliability for t_B (ICC = 0.851, p < 0.001; ICC = 0.798, $p = 0.001$, respectively).

Discriminant validity (Studies 1 and 2)

Table 3 shows AUC values from ROC curves (shown in Additional file 1). Pencil-and-paper CTT AUC values were compared with AUC values obtained for each of the VR adaptations (i.e., DOME-CTT, HMD-CTT). For each VR adaptation, the relative difference [%] in AUC from the pencil-and-paper AUC is shown. Comparisons were made separately for Trails A and Trails B.

The data indicate that all CTT versions have relatively high discriminant validity (AUC ≥ 0.70; $p \leq 0.05$). AUCs are largely comparable, though slightly reduced for the VR adaptations.

Comparing completion times between DOME-CTT and HMD-CTT

The completion times in Tables 1 and 2 suggest that t_A and t_B are higher (i.e., longer) for DOME-CTT relative to HMD-CTT. None of the participants completed both DOME-CTT and HMD-CTT testing, and the present work was not designed to compare between the two VR platforms. However, we conducted analyses to address this question based on the existing evidence. The methodology and the results detailed in the supplemental material (Additional file 1) suggest shorter completion times for HMD-CTT as compared to the DOME-CTT (Tables 1 and 2).

Discussion

In this report, we describe the development and initial validation of VR-based adaptations of the Color Trails Test (CTT) [27], a traditional pencil-and-paper test of attention and processing speed, using two different types of VR systems. The first VR-based system involves projection of visual stimuli on the walls of a large, dome-shaped room (akin to a cave, monoscopic projection).

Table 3 AUC values from ROC curves

	Pencil-and-Paper-CTT		**DOME-CTT**	
	Trails A (t_A)	**Trails B (t_B)**	**Trails A (t_A)**	**Trails B (t_B)**
YA vs. MA	0.88**	0.82**	0.88** [0%]	0.81** [− 2%]
	Pencil-and-Paper-CTT		**HMD-CTT**	
YA vs. MA	0.73*	0.77*	0.70* [− 5%]	0.71* [− 8%]
YA vs. OLD	0.95*	0.95*	0.92* [-4%]	0.92* [− 4%]
MA vs. OLD	0.83**	0.80**	0.75** [-10%]	0.80** [0%]

*p = .05; ** p < .0003; []—relative difference of AUC for VR adaptation as compared with Pencil-and Paper CTT

AUC area under the curve, *ROC* receiver operating characteristic curves, *YA* young adults, *MA* middle aged adults, *OLD* Elderly adults

The second VR-based system is a low-cost head-mount VR device (HMD) worn by the participant and suitable for home-based assessment. Adherence and usability of both VR-based adaptations proved to be relatively good, with only two participants (~ 1.5%) not completing the VR tasks. Participants only rarely complained about the difficulty of completing the VR tasks, though there were no such complaints for the pencil-and-paper version. Our discussion integrates the results from two studies, each of which evaluated one of the VR-based adaptations.

Construct validity

Our results suggest that the new VR-based adaptations and gold standard pencil-and-paper version share similar psychometric properties (e.g., longer completion time for B vs. A). Coupled with the relatively high correlations between corresponding parts (~ 0.7; Figs. 5 and 7), this suggests that the VR and the pencil-and-paper tests measure the same cognitive constructs (e.g., sustained, divided attention). By comparison, in a cross-validation study of the TMT and CTT, Dugbartey and colleagues reported lower correlation values of 0.35 for Trails A and 0.45 for Trails B [35].

Notably, construct validity correlations for YA participants on the Trails A portion of the test were not significant. We attribute this result to a ceiling effect in that t_A values approached the shortest completion times technically possible on all three CTT tests.

Completion time: format effects

Trails A and Trails B completion times were significantly longer for the VR-based adaptations compared with the pencil-and-paper CTT, possibly reflecting a larger dynamic range of performance for the VR versions (e.g., even if only attributable to the larger spatial area covered by the VR task as compared to one page distribution of the targets in the pencil-and-paper CTT), greater task difficulty [36] and/or the cognitive-motor interactions relevant to the VR versions but not the original pencil-and-paper test.

Perceptual factors must also be considered. While the participant has an egocentric viewpoint for both pencil-and-paper and VR-based versions [37], s/he likely has different perceptions of the candidate actions available to perform the task (i.e., different perception of affordances) [38–40]. Presumably, during the pencil-and-paper test, visual scanning is mainly by saccadic eye movements and short-distance visual pursuits. In contrast, the VR versions require head "gaze" (i.e., motor programs for the neck and upper trunk muscles to execute head rotations mainly around the yaw and pitch axes) combined with longer ocular pursuits and with saccades for operational visual scanning. Further, in the pencil-and-paper CTT, motor activity of the hand is limited to drawing short lines between the printed circles. However, in the VR-CTT versions, larger arm reaching movements are required as well as postural adjustments and occasional stepping (multi directional). It is conceivable that prior to a given task and during the short practice levels, the participant 'tunes' his/her perception of the affordance related to the task, which includes the more complex integration required for the various actions associated with the VR-CTT versions.

Finally, we speculate that the three-dimensional target layout within a black space with perceived infinite boundaries, i.e., unknown physical limits of the VR-CTT versions as compared with the finite boundaries of the pencil-and-paper CTT version, contributes to longer test execution times. Specifically, the participant may have difficulty with movement scaling in the absence of physical reference boundaries on the VR-based versions. This speculation can be tested in future studies by the use of a VR version that includes virtual physical boundaries (e.g., target balls floating in a room).

Among the two VR adaptations, completion times for the DOME-CTT were longer than those for the

HMD-CTT (compare Tables 1 and 2, Figs. 5 and 7; post-hoc analyses comparing across studies—Additional file 1). One possible account for this finding relates to different levels of visual immersion between the tests. The HMD-CTT provides no visual feedback from the arms, and the participant's subjective experience consists solely of moving the avatar (red ball) within the VR environment. In contrast, during the DOME-CTT, the participant sees his/her hand holding the wand-like stick in addition to the virtual avatar as s/he makes reaching movements toward the target balls. The latter configuration may complicate sensorimotor integration given the two parallel, relevant sensory input streams (physical hand, virtual avatar). A potential contributor to this complication is that participants can see their arms in full stereoscopic vision but the balls only in monoscopic projection.

Subramanian and Levin reported superior motor performance (reaching movements) among healthy adults in a large-scale screen-based VR system as compared to an HMD-based system [41], apparently at odds with the present findings (i.e., slower movements for DOME-CTT as compared with HMD-CTT). Likely technical-methodological differences between the studies account for the disparity. For example, the HMD field of view was smaller (~50° vs. ~100°) in the Subramanian and Levin study, and the type of task (reaching vs. consecutive trails making following rules) was markedly different.

Error performance

Participants made significantly more errors (i.e., touching the wrong ball) on the VR-based versions as compared to the pencil-and-paper CTT (Figs. 4 and 6). For example, only about 14% of all pencil-and-paper test levels completed in Study 2 (Trials A and B across all three cohorts) had at least one error (most often one error). The error rate was 2.5 times higher for HMD-CTT test levels (~35%).

This pattern of results may seem paradoxical, as with longer completion times on the VR-based tests, fewer errors should occur, but our data reflect the opposite. Indeed we believe that as the VR tasks are more demanding then the corresponding pencil-and-paper tasks, the cognitive processes classically associated with the CTT paradigm might be compromised, leading to more errors [42–45]. Specifically, we posit that the VR-based tasks make much greater demands on motor planning and execution (see below), visual scanning and spatial orientation, and involve higher perceptual and/or cognitive load. Apparently, this load differentially affected elderly as compared to YA, as evidenced by the significantly higher error rates for the OLD group.

Cognitive-motor interactions

Our qualitative analyses clearly demonstrate that when shifting from a cognitive task primarily involving sustained attention (Trails A) to one that primarily involving divided attention (Trails B), upper-limb motor behavior changes (Fig. 8). Previous research has employed VR to evaluate cognitive-motor interactions mainly in the context of locomotion. Most of the studies have reported clinical benefits related to cognitive-motor interactions associated with immersion in a VR environment [46–50].

In the current study, we began exploring the effect of divided attention (operationalized as HMD-CTT Trails B performance) on the planning and execution of upper-limb reaching movements. The well-documented single-peak velocity profile typical of ballistic movements [51–53] appears to govern the hand trajectories generated during HMD-CTT Trails A, a sustained attention task, but not during Trails B. In Trails B, a divided attention task, trajectories are characterized by an initial slow increase in the velocity profile, probably reflecting neural processes more related to executive function and less to motor execution. Potential age effects (e.g., less symmetric peak, slower overall velocity) are apparent in comparing the velocity profile across age groups (Fig. 8).

Follow up studies will focus on developing reliable quantification methods and metrics to assess these cognitive-motor interaction effects. Notably, comparing the velocity profiles generated during a three-dimensional (3D) VR-based task to a classical two-dimensional (2D) task as the 'gold standard' [54] is suboptimal, mainly due to the absence of a reliable theoretical model for three-dimensional hand-reaching movements. Thus, new referencing methodologies like sampling single target-to-target trajectories should be included as part of future versions and analyses of VR-based CTT tasks like those used here.

Discriminant validity

The VR-based tests were largely comparable though not superior to the pencil-and-paper CTT in terms of distinguishing individuals of different age-groups on the basis of CTT completion time (t_A and t_B, Table 3). Further, both the traditional and VR-based versions demonstrated relatively high discriminant validity as reflected by high AUC values. These observations are consistent with the strong correlations for completion times between the VR-based and original CTT for each age group (with the exception of t_A in YA) as well as when combining participants across age groups (black dashed lines in Figs. 5 and 7).

Comparability in discriminant validity between VR-based and the gold standard CTT for completion times is encouraging. However, VR-based testing affords additional metrics that may be better at differentiating among age groups as well as between healthy and cognitively impaired individuals. Indeed, VR facilitates the development of new parameters of greater relevance to daily living (i.e., more ecologically valid) in that they better capture complex, integrated behaviors. Thus, we speculate that using such VR-based parameterization of multimodal function (e.g., hand-gaze coordination combined with hand trajectories) will provide superior discriminant validity.

Test retest reliability

For a retest period of ~ 12 weeks, the VR-based CTT adaptations showed moderate reliability (intraclass correlation of ~ 0.6), while the pencil-and-paper version showed generally better reliability. The superior reliability of the original CTT for this retest interval may be attributable to the greater familiarity of the pencil-and-paper format, which may have led to a larger learning effect upon retest and consequently poorer reliability for the VR-based versions (see [55]). We also acknowledge that some middle-aged participants had engaged in a cognitive training protocol during the 12-week interval (see Additional file 1: Table S1) which may compromise test-retest evaluations.

However, for a retest period of ~ 2 weeks, both the HMD-CTT and the original CTT showed good reliability (intraclass correlation of ≥ 0.75), with the VR-based adaptation showing superior reliability for Trails B. Our results are consistent with findings that shorter retest intervals yield higher reliability coefficients [56]. As there does not appear to be a clear convention for the ideal test–retest interval [57], our data reporting reasonable reliability for both intervals is relevant and informative. Still, given that our sample sizes were small, the findings should be replicated in larger studies.

Limitations

This study had several notable methodological/technical limitations. Some concepts could not be directly translated from the pencil-and-paper CTT to the VR-CTT versions. For example, we provided positive feedback upon reaching the correct target ball in the VR versions, while only negative feedback is provided pencil-and paper version (i.e., when drawing a line to the wrong circle). Our decision to provide positive feedback in the VR versions was to assure the participant that the target had ben successfully reached. In addition, to familiarize the user with the VR environment, more practice sessions were performed as compared to the pencil-and-paper versions, which might introduce learning effects.

As detailed in Additional file 1: Table S1, the 147 participants in this study performed the CTT tasks while participating in different larger protocols. This could potentially affect, e.g., the test–retest reliability results. However, our post-hoc analyses did not show substantial effect (see *Results* & Additional file 1).

Some limitations are related to the VR media used. For example, visual acuity of the participant is more critical for performance of the HMD-CTT versions than for the pencil-and-paper version, in which the paper remains at a constant, comfortable distance at which all potential targets are visible. The 3D HMD-CTT is fundamentally different in this respect, as the potential targets are located at a variety of virtual depths.

Future directions

VR technologies may enable us to enrich the current VR-based versions of the CTT to further enhance ecological relevance, mainly in the sense of engaging more modalities, and inter-modalities interactions.

The challenge will then be how to leverage multimodal measures to understand such real-world processes as cognitive-motor interference during multi-tasking and ultimately assess function in a predictive or clinically meaningful way. In particular, we hope to achieve superior discriminant validity for patient cohorts and the ability to predict risks associated with impaired cognitive-motor interactions [58–61], such as the risk of falls in the elderly and in neurological patients [62, 63].

Finally, we envision developing adaptations of additional neuropsychological tests, with different core construct (i.e., than the CTT) for application in an immersive VR environment.

Conclusions

In sum, the present study describes the development and validation of large-scale (DOME-CTT) and head-mount (HMD-CTT) VR adaptations of the classic pencil-and-paper Color Trails Test (CTT) and provides key validation data, including construct validity relative to the original test, discriminant validity among age groups, and test–retest reliability at two different retest intervals. Critically, this work demonstrates the feasibility and viability of converting a neuropsychological test from two-dimensional pencil-and-paper to three-dimensional VR based format while preserving core features of the task and assessing the same cognitive functions. Our novel findings on the relationship between classical cognitive performance and upper-limb motor planning and execution may lead to new analysis methods for other more ecological VR-based neuropsychological tests that incorporate cognitive-motor interactions.

Abbreviations

ANOVA: Analysis of variance; AUC: Area under the curve; CTT: Color trails test; FOV: Field of view; HMD: Head mounted device; ICC: Intraclass correlation coefficients; MA: Middle-aged adults; MET: Multiple errands test; ROC: Receiver operating characteristic; SD: Standard deviation; TMT: Trail making test; VET: Virtual errands test; VR: Virtual reality; YA: Young adults.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12984-021-00849-9.

Additional file 1: Additional information on Methods and Results.

Additional file 2: Video demo of DOME-CTT.

Additional file 3: Video demo of HMD-CTT.

Acknowledgements

The authors wish to thank Ms. Lotem Kribus Shmiel and Mr. Evyatar Arad for assistance in designing the DOME-CTT, Prof. Yuval Heled, Dr. Ran Yanovich, Ms. Hadar Sharon, Ms. Anastsaia Tkachov, Mr. Hani Baransi and Mr. Yuval Ketko, for their collaboration in data collection, and Ms. Hila Iny for technical assistance.

Authors' contributions

MP—Designed the study, supervised its execution, analyzed the data and drafted the manuscript; OBG—participated in designing the virtual reality environments, in analyzing the data and in data collection, GMD– contributed to drafting the manuscript and in data collection; AG—contributed to data analyses; YB—designed and supervised the programing of the virtual reality environments; MC—contributed to data collection and handling; SKN—contributed to data collection and analyses; GZ supervised ethical aspects of the study; MSB—participated in supervising the study, recruitment of participants and provided iterative critical commenting on the drafts of the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported in part by the Sheba Medical Center Research Fund and by a donation from the Nissim family.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All participants in the study provided their informed consent to participate. All study procedures were approved by the Institutional Review Board of Sheba Medical Center.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center of Advanced Technologies in Rehabilitation, Sheba Medical Center, Ramat Gan, Israel. [2]Department of Physiology and Pharmacology, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel. [4]Joseph Sagol Neuroscience Center, Sheba Medical Center, Ramat Gan, Israel. [5]Department of Neurological Rehabilitation, Sheba Medical Center, Ramat Gan, Israel. [6]Deartment of Rehabilitation, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [7]Department of Psychiatry, The Icahn School of Medicine At Mount Sinai, New York, NY, USA.

Received: 23 April 2020 Accepted: 12 March 2021
Published online: 17 May 2021

References

1. Chan RC, Shum D, Toulopoulou T, Chen EY. Assessment of executive functions: review of instruments and identification of critical issues. Arch Clin Neuropsychol. 2008;23(2):201–16.
2. Diamond A. Executive functions. Annu Rev Psychol. 2013;64:135–68.
3. Burgess PW, Alderman N, Evans J, Emslie H, Wilson BA. The ecological validity of tests of executive function. J Int Neuropsychol Soc. 1998;4(6):547–58.
4. Bottari C, Dassa C, Rainville C, Dutil E. The factorial validity and internal consistency of the Instrumental activities of daily living profile in individuals with a traumatic brain injury. Neuropsychol Rehabil. 2009;19(2):177–207.
5. Manchester D, Priestley N, Jackson H. The assessment of executive functions: coming out of the office. Brain Inj. 2004;18(11):1067–81.
6. Sbordone RJ. Ecological validity of neuropsychological testing: critical issues. Neuropsychol Handbook. 2008;367:394.
7. Burgess PW, Alderman N, Forbes C, Costello A, Laure MC, Dawson DR, et al. The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology. J Int Neuropsychol Soc. 2006;12(2):194–209.
8. Goldstein G. Functional considerations in neuropsychology. 1996.
9. Rabin LA, Burton LA, Barr WB. Utilization rates of ecologically oriented instruments among clinical neuropsychologists. Clin Neuropsychol. 2007;21(5):727–43.
10. Shallice T, Burgess PW. Deficits in strategy application following frontal lobe damage in man. Brain. 1991;114(2):727–41.
11. Parsons TD, Carlew AR, Magtoto J, Stonecipher K. The potential of function-led virtual environments for ecologically valid measures of executive function in experimental and clinical neuropsychology. Neuropsychol Rehabil. 2017;27(5):777–807.
12. Logie RH, Trawley S, Law A. Multitasking: multiple, domain-specific cognitive functions in a virtual environment. Mem Cognit. 2011;39(8):1561–74.
13. McGeorge P, Phillips LH, Crawford JR, Garden SE, Sala SD, Milne AB, et al. Using virtual environments in the assessment of executive dysfunction. Presence . 2001;10(4):375–83.
14. Claessen MH, Visser-Meily JM, de Rooij NK, Postma A, van der Ham IJ. A direct comparison of real-world and virtual navigation performance in chronic stroke patients. J Int Neuropsychol Soc. 2016;22(4):467–77.
15. Kimura K, Reichert JF, Olson A, Pouya OR, Wang X, Moussavi Z, et al. Orientation in virtual reality does not fully measure up to the real-world. Sci Rep. 2017;7(1):1–8.
16. Elkind JS, Rubin E, Rosenthal S, Skoff B, Prather P. A simulated reality scenario compared with the computerized Wisconsin Card Sorting Test: an analysis of preliminary results. Cyberpsychol Behav. 2001;4(4):489–96.
17. Josman N, Kizony R, Hof E, Goldenberg K, Weiss PL, Klinger E. Using the virtual action planning-supermarket for evaluating executive functions in people with stroke. J Stroke Cerebrovasc Dis. 2014;23(5):879–87.
18. Nir-Hadad SY, Weiss PL, Waizman A, Schwartz N, Kizony R. A virtual shopping task for the assessment of executive functions: validity for people with stroke. Neuropsychol Rehabil. 2017;27(5):808–33.
19. Rizzo AS, Koenig ST. Is clinical virtual reality ready for primetime? Neuropsychology. 2017;31(8):877–99.
20. Davison SMC, Deeprose C, Terbeck S. A comparison of immersive virtual reality with traditional neuropsychological measures in the assessment of executive functions. Acta Neuropsychiatr. 2017. https://doi.org/10.1017/neu.2017.14.
21. Parsons TD, Barnett MD. Virtual apartment stroop task: Comparison with computerized and traditional stroop tasks. J Neurosci Methods. 2018;309:35–40.
22. Ouellet E, Boller B, Corriveau-Lecavalier N, Cloutier S, Belleville S. The Virtual Shop: A new immersive virtual reality environment and scenario for the assessment of everyday memory. J Neurosci Methods. 2018;303:126–35.
23. Parsons TD. Ecological validity in virtual reality-based neuropsychological assessment. In: Mehdi Khosrow-Pour DBA, editor. Encyclopedia of

information science and technology. 3rd ed. Hershey: IGI Global; 2015. p. 1006–15.

24. Parsons TD. Neuropsychological assessment 3.0. clinical neuropsychology and technology: what's new and how we can use it. Cham: Springer International Publishing; 2016. p. 65–96.
25. Reitan RM. Validity of the trail making test as an indicator of organic brain damage. Percept Mot Skills. 1958;8(3):271–6.
26. Reitan RM, Wolfson D. Category test and trail making test as measures of frontal lobe functions. Clin Neuropsychol. 1995;9(1):50–6.
27. D'Elia L, Satz P, Uchiyama CL, White T. Color Trails Test: CTT: psychological assessment resources Odessa, FL; 1996.
28. D'Elia L, Satz P. Color trails test: psychological assessment resources; 2000.
29. Levene H. Robust tests for equality of variances. Contributions to probability and statistics Essays in honor of Harold Hotelling. 1961:279–92.
30. Best J. How virtual reality is changing medical practice:"Doctors want to use this to give better patient outcomes." BMJ. 2019. https://doi.org/10.1136/bmj.k5419.
31. Le Chénéchal M, Goldman JC. HTC Vive Pro time performance benchmark for scientific research. International Conference on Artificial Reality and Telexistence (ICAT)-Eurographics Symposium on Virtual Environments (EGVE); Limassol, Cyprus2018.
32. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.
33. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284.
34. Jutten RJ, Harrison J, Kjoe PRLM, Opmeer EM, Schoonenboom NS, de Jong FJ, et al. A novel cognitive-functional composite measure to detect changes in early Alzheimer's disease: test–retest reliability and feasibility. Alzheimers Dement. 2018;10:153–60.
35. Dugbartey AT, Townes BD, Mahurin RK. Equivalence of the color trails test and trail making test in nonnative English-speakers. Arch Clin Neuropsychol. 2000;15(5):425–31.
36. Neguț A, Matu S-A, Sava FA, David D. Task difficulty of virtual reality-based assessment tools compared to classical paper-and-pencil or computerized measures: a meta-analytic approach. Comput Hum Behav. 2016;54:414–24.
37. Chellali A, Milleville-Pennel I, Dumas C. Influence of contextual objects on spatial interactions and viewpoints sharing in virtual environments. Virtual Reality. 2013;17(1):1–15.
38. Fajen BR. Guiding locomotion in complex, dynamic environments. Front Behav Neurosci. 2013;7:85.
39. Fajen BR. Affordance perception and the visual control of locomotion. In: Steinicke F, Visell Y, Campos J, Lécuyer A, editors. Human walking in virtual environments. New York: Springer; 2013. p. 79–98.
40. Fajen BR, Matthis JS. Direct perception of action-scaled affordances: The shrinking gap problem. J Exp Psychol Hum Percept Perform. 2011;37(5):1442.
41. Subramanian SK, Levin MF. Viewing medium affects arm motor performance in 3D virtual environments. J Neuroeng Rehabil. 2011;8(1):36.
42. Dinstein I, Gardner JL, Jazayeri M, Heeger DJ. Executed and observed movements have different distributed representations in human aIPS. J Neurosci. 2008;28(44):11231–9.
43. Kannape OA, Barré A, Aminian K, Blanke O. Cognitive loading affects motor awareness and movement kinematics but not locomotor trajectories during goal-directed walking in a virtual reality environment. PLoS ONE. 2014;9(1):e85560.
44. Makransky G, Terkildsen TS, Mayer RE. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. Learn Instr. 2019;60:225–36.
45. Olk B, Zielinski D, Kopper R. Effects of perceptual load in visual search in immersive virtual reality. J Vis. 2015;15(12):1064.
46. Cano Porras D, Sharon H, Inzelberg R, Ziv-Ner Y, Zeilig G, Plotnik M. Advanced virtual reality-based rehabilitation of balance and gait in clinical practice. Ther Adv Chronic Dis. 2019;10:2040622319868379.
47. Dockx K, Bekkers EM, Van den Bergh V, Ginis P, Rochester L, Hausdorff JM, et al. Virtual reality for rehabilitation in Parkinson's disease. Cochrane Database Syst Rev. 2016. https://doi.org/10.1002/14651858.CD010760.pub2.
48. Laver KE, Lange B, George S, Deutsch JE, Saposnik G, Crotty M. Virtual reality for stroke rehabilitation. Cochrane Database Syst Rev. 2017. https://doi.org/10.1002/14651858.CD008349.pub4.
49. Mirelman A, Rochester L, Maidan I, Del Din S, Alcock L, Nieuwhof F, et al. Addition of a non-immersive virtual reality component to treadmill training to reduce fall risk in older adults (V-TIME): a randomised controlled trial. The Lancet. 2016;388(10050):1170–82.
50. Pereira VAI, Polastri PF, Simieli L, Rietdyk S, Imaizumi LFI, Moretto GF, et al. Parkinson's patients delay fixations when circumventing an obstacle and performing a dual cognitive task. Gait Posture. 2019;73:291–8.
51. Flash T, Hogan N. The coordination of arm movements: an experimentally confirmed mathematical model. J Neurosci. 1985;5(7):1688–703.
52. Kawato M. Internal models for motor control and trajectory planning. Curr Opin Neurobiol. 1999;9(6):718–27.
53. Uno Y, Kawato M, Suzuki R. Formation and control of optimal trajectory in human multijoint arm movement. Biol Cybern. 1989;61(2):89–101.
54. Gal OB, Doniger GM, Cohen M, Bahat Y, Plotnik M. Cognitive-motor interaction during virtual reality trail making. In: 2019 International Conference on Virtual Rehabilitation (ICVR). IEEE; 2019. pp. 1–6.
55. Heilbronner RL, Sweet JJ, Attix DK, Krull KR, Henry GK, Hart RP. Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: the utility and challenges of repeat test administrations in clinical and forensic contexts. Clin Neuropsychol. 2010;24(8):1267–78.
56. Duff K. Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. Arch Clin Neuropsychol. 2012;27(3):248–61.
57. Llorente AM, Voigt RG, Williams J, Frailey JK, Satz P, D'Elia LF. Children's Color Trails Test 1 & 2: test–retest reliability and factorial validity. Clin Neuropsychol. 2009;23(4):645–60.
58. Berard J, Fung J, Lamontagne A. Impact of aging on visual reweighting during locomotion. Clin Neurophysiol. 2012;123(7):1422–8.
59. Demanze Laurence B, Michel L. The fall in older adults: physical and cognitive problems. Curr Aging Sci. 2017;10(3):185–200.
60. Lord SR, McLean D, Stathers G. Physiological factors associated with injurious falls in older people living in the community. Gerontology. 1992;38(6):338–46.
61. Montero-Odasso M, Almeida QJ, Bherer L, Burhan AM, Camicioli R, Doyon J, et al. Consensus on shared measures of mobility and cognition: from the Canadian Consortium on Neurodegeneration in Aging (CCNA). The J Gerontol Series A. 2019;74(6):897–909.
62. de Rooij IJ, van de Port IG, Meijer J-WG. Effect of virtual reality training on balance and gait ability in patients with stroke: systematic review and meta-analysis. Phys Ther. 2016;96(12):1905–18.
63. Porras DC, Siemonsma P, Inzelberg R, Zeilig G, Plotnik M. Advantages of virtual reality in the rehabilitation of balance and gait: systematic review. Neurology. 2018;90(22):1017–25.

Publisher's Note