Computers in Human Behavior 54 (2016) 414–424

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Literature review

Task difficulty of virtual reality-based assessment tools compared to classical paper-and-pencil or computerized measures: A meta-analytic approach

Alexandra Neguţ [a, *], Silviu-Andrei Matu [b], Florin Alin Sava [c], Daniel David [b, d]

[a] Evidence-based Assessment and Psychological Interventions Doctoral School, The International Institute for the Advanced Studies of Psychotherapy and Applied Mental Health, Babeş-Bolyai University, No. 37, Republicii Street 400015, Cluj-Napoca, Cluj, Romania
[b] Department of Clinical Psychology and Psychotherapy, Babeş-Bolyai University, No. 37, Republicii Street 400015, Cluj-Napoca, Cluj, Romania
[c] Department of Psychology, West University of Timişoara, No. 4, Vasile Pârvan Bulevard 300223, Timişoara, Timiş, Romania
[d] Icahn School of Medicine at Mount Sinai, New York, USA

ARTICLE INFO

Article history:
Received 28 May 2015
Received in revised form
25 July 2015
Accepted 23 August 2015
Available online 3 September 2015

Keywords:
Virtual reality
Neuropsychological assessment
Task difficulty

ABSTRACT

Virtual reality-based assessment tools arise as a promising alternative for classic neuropsychological assessment with an increased level of ecological validity. Because virtual reality cognitive measures recreate tasks that resemble with the demands from the real world it is assumed that they require additional cognitive resources and are more difficult than classical paper-and-pencil or computerized measures. Although research has focused on comparing the performance obtained on virtual reality-based measures with classical paper-and-pencil or computerized measures, no meta-analysis has been conducted on this topic. Thirteen studies met our inclusion criteria: assessed any cognitive process using virtual reality and analogous classical or computerized assessment tools of the same process. Based on a random effects model, the results indicated a moderate effect size in favor of classical and computerized tests ($g = -0.77$) revealing an increased task difficulty in virtual reality. Overall, results from the current meta-analysis point out that cognitive performance obtained in virtual reality is poorer than the one in classical or computerized assessment which might suggest that tasks embedded in virtual reality have an increased level of complexity and difficulty and require additional cognitive resources.

Contents

* Corresponding author.
E-mail addresses: alexandra_negut@yahoo.com (A. Neguţ), silviu.matu@ubbcluj.ro (S.-A. Matu), afsava@gmail.com (F.A. Sava), danieldavid@psychology.ro (D. David).

http://dx.doi.org/10.1016/j.chb.2015.08.029
0747-5632/

1. Introduction

Virtual reality consists of a human–computer interface that is based on an interactive and advanced computer technology. By using a wide ranges of technological tools like head-mounted displays (HMDs) for the visual input, trackers, headphones for acoustic input, video capture systems, data gloves or joysticks a 3D environment is generated (Gamberini, 2000; Ku et al., 2003; Parsons, 2012; Rand et al., 2005; Schultheis, Himelstein, & Rizzo, 2002). The virtual environment generated by the technological tools is a computerized representation of the real world. The person is immersed in the virtual environment and is able to interact with it. Immersion generates a sense of presence in the world, as if he is actually present in the computer-generated world (Elkind, Rubin, Rosenthal, Skoff, & Prather, 2001; Ku et al., 2003; Lalonde, Henry, Drouin-Germain, Nolin, & Beauchamp, 2013; Rheingold, 1991).

1.1. Main approaches to cognitive assessment

Neuropsychological assessment is considered an applied science that focuses on the evaluation of specific activities in the central nervous system that are associated with observable behaviors (Lezak, 1995). Classic paper-and-pencil psychometrics, as well as computer-based assessment instruments, represents the current standard assessment tools used in neuropsychological evaluation (Podell, DeFina, Barrett, McCullen, & Goldberg, 2003). They consist of a certain amount of stimuli delivered to the subjects in a highly systematic and controlled environment via written paper or a computer screen. A recent study (Holzinger et al., 2011) shows that when taking into account the performance of medical professionals in a real-life setting on visual productivity between classical paper presentation and computerized screens no differences emerge, but the use of paper presentation is more preferred (Holzinger et al., 2011). Also, scoring and test interpretation are conducted either by a trained practitioner or automatically by the computer (Bauer et al., 2012; Podell et al., 2003). However, because the task characteristics associated with classical and computerized assessment do not replicate the complexity and challenges found in everyday life, their predictive power for real life performance in is limited (Armstrong et al., 2013; Elkind, 1998; Rizzo, Schultheis, Kerns, & Mateer, 2004; Schultheis et al., 2002). Considering these drawbacks, there is need to develop other assessment instruments with increased ecological validity (Alvarez & Emory, 2006; Elkind, 1998; Schultheis et al., 2002).

Virtual reality neuropsychological assessment might represent an efficient alternative to classical or computerized tests, given that it provides a higher level of ecological validity. Ecological validity implies a close link between the challenges imposed by the assessment procedures and the challenges that the subject has to confront in real life situations (Wasserman & Bracken, 2003). Virtual reality-based tests can increase the ecological validity of the assessment because they simulate real-life stressors and replicate the challenges and distractors found in day to day situations (Pugnetti et al., 1998; Rizzo et al., 2004; Schultheis et al., 2002). In addition, they may have potential to predict real life functioning due to the characteristics of test administration and assessment context (Elkind, 1998; Rizzo et al., 2006).

Virtual reality instruments are used for the neuropsychological assessment of executive functions, attention, and impulsivity, cognitive and motor inhibition (Adams, Finn, Moes, Flannery, & Rizzo, 2009; Elkind, 1998; Henry, Joyal, & Nolin, 2012; Ku et al., 2003; Parsons, Courtney, Arizmendi, & Dawson, 2011), memory and learning (Gamberini, 2000; Matheis et al., 2007; Parsons & Rizzo, 2008; Pugnetti et al., 1998), spatial abilities (Parsons et al., 2004), and visuospatial neglect (Broeren, Samuelsson, Stibrant-Sunnerhagen, Blomstrand, & Rydmark, 2007). Results of these studies support the use of virtual reality scenarios in neuropsychological assessment because they discriminate between healthy and clinical populations and their accuracy is similar to classical tests. Furthermore, results show a good equivalence between the performance obtained in the virtual world and in the real world (Rand, Basha-Abu Rukan, Weiss, & Katz, 2009; Sorita et al., 2013).

1.2. Overview of the current study

Due to the high similarity with the real world demands, it seems that virtual reality-based assessment has an increased task difficulty and triggers more cognitive resources than classical or computerized psychometrics (Elkind, 1998; Gamberini, 2000). Further on, the visual complexity of an interface influences the overall performance (Stickel, Ebner, & Holzinger, 2010) and virtual reality has an increased visual complexity compared to classical or computerized assessment, because it recreates a real environment. Overall, virtual reality scenarios replicate more accurately the complexity of real world situations which can lead to poorer performance on cognitive tasks conducted in virtual environments than on classical or computerized measures (Armstrong et al., 2013; Broeren et al., 2007; Gamberini, 2000; Parsons & Courtney, 2014; Parsons, Courtney, & Dawson, 2013). However, despite the fact that previous research has provided a useful database on the topic of virtual-reality based neuropsychological assessment and a reasonable number of theoretical reviews provide useful information about the core aspects and advantages of virtual reality assessment (Elkind, 1998; Myers & Bierig, 2000; Riva, 1998; Rizzo et al., 1999) no meta-analysis has been conducted in order to investigate the task difficulty hypotheses of virtual reality assessment tools in comparison to classical or computerized instruments. Although the current findings in the

literature point out that virtual reality measures are more complex and difficult because they replicate conditions similar to everyday life, and as a consequence performance obtained on virtual reality tests is usually poorer than the performance on classical measures, there is need to conduct a meta-analysis to make sense of a collective body of research findings without bias. Meta-analysis can overcome the drawbacks of narrative reviews such as selective bias of studies, and offers a common yardstick to compare across studies by converting inferential statistics to an effect size. Nevertheless, giving that virtual reality neuropsychological assessment techniques are spreading in both scientific and clinical communities, and their potential benefits over classical and computerized measures, a meta-analysis could help clarify important issues regarding their task difficulty and complexity.

Therefore, the current meta-analysis sought to examine the following objectives:

1. To examine differences in performance between classical or computerized measures and virtual reality-based measures of cognitive processes;
2. To investigate potential moderators of the results.

1.3. Potential theoretical moderator variables

1.3.1. Demographic variables

We consider participants' mean age and percentage of male participants as potential moderators of the overall effect. First, age can moderate the strength of the effect. Previous exposure to technology yields an impact over its acceptance (Holzinger, Searle, & Wernbacher, 2011) and children may be more attracted and more familiarized to technology than adults are. Also, Next, young adults may be more motivated to complete and succeed on tests. Also, it is well known the tendency of cognitive processes to decline among samples of older adults (Urbina, 2004). Further on, gender may influence the effect due to a superiority of male participants than female on spatial navigation tasks (Parsons et al., 2004; Voyer, Voyer, & Bryden, 1995). Previous research conducted on virtual reality-based assessment has never considered age or gender as moderator variables.

1.3.2. Clinical status of the sample

We anticipate that the effects for the comparison of virtual reality-based measure with paper-and-pencil and computer-based measures will be larger in case of healthy participants than for clinical participants, because cognitive impairment associated with clinical condition will decrease the impact of task difficulty. In other words, because clinically impaired participants will perform worse than controls on both virtual reality-based measures and classical or computerized measures, the difference in results between types of assessment instruments will be smaller for the clinical populations (Elkind et al., 2001; Gamberini, 2000).

1.3.3. Type of control measurement instrument

It is common in psychological testing to program a classical paper-and-pencil test for computer administration. In this case, the test becomes a computerized test of the same psychological construct. Yet, the computerized test is a new and different measurement instrument with different psychometric properties (Bauer et al., 2012). Due to such theoretical and methodological considerations we investigated the moderating effect of type of measurement instrument.

1.3.4. Task performance indicator

We classified task performance indicator in two main clusters: (1) based on errors, such as correct or incorrect responses, and (2) based on time, such as reaction time. It is possible that time-based measures require different cognitive resources than error-based measures and this distinction may be better expressed via virtual reality-based assessment.

2. Method

2.1. Literature search

In order to identify potentially relevant studies, a systematic literature search on virtual reality assessment has been conducted using “virtual reality”, “cogn* assessment”, “memory”, “executive funct*”, and “attention” as search terms in Medline, PsychInfo and ScienceDirect databases, up to November 2014. Furthermore, the list of references of empirical articles and reviews on this topic were screened in order to detect other studies that did not appear in the electronic search.

2.2. Studies selection

The following criteria were used for the inclusion of studies in the meta-analysis: (a) assessed any cognitive process using virtual reality and analogous classical or computerized assessment tools of the same cognitive process; (b) provided sufficient data to compute effect sizes; (c) were English-based publications.

The initial search procedure revealed 146 records. Thirty-three additional records were identified through other sources (see Fig. 1). After removing 16 duplicates, 163 potential abstracts were inspected. We excluded dissertations, publications in other languages than English, and studies that were not focused on virtual reality and neuropsychological assessment. A total of 115 potential articles were analyzed in detail based on their full text. Studies that used computer devices but did not provide full immersion via HMDs or gesture-based video-capture systems have been excluded. Thirteen studies met the inclusion criteria and were included in the meta-analysis.

2.3. Data coding

The following variables were coded: study identification data, participants' mean age, percentage of male participants, number of participants per condition, clinical status of the sample, type of clinical condition, type of control measurement instrument, task performance indicator, type of cognitive process, type of virtual reality platform.

Outcome measures were classified into three categories based on the cognitive process assessed, and subsequent cognitive assessment scales: executive functions, memory, and other neurocognitive measures. Only these measures were available for analysis from the studies that met the inclusion criteria in the meta-analysis.

Executive functions measures included general measures of executive functioning, as well as attention indexes/measures, and impulsivity/inhibition measures.

The memory measures outcome included memory and learning processes (e.g., incidental memory, target recall, target recognition, object recognition).

The measures grouped under the final category of outcomes (other neurocognitive measures) included measures of spatial rotation and measures of visuospatial neglect.

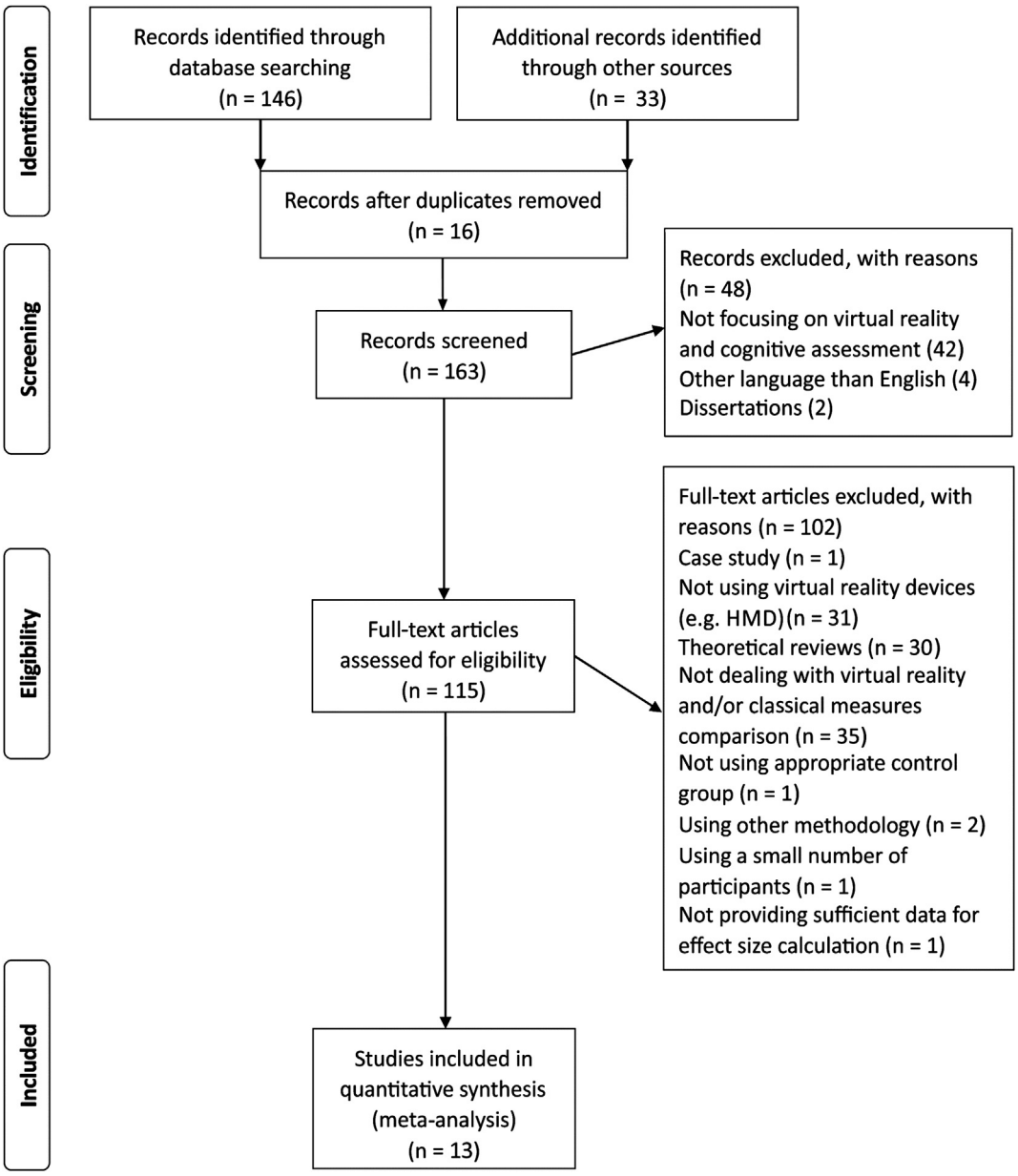

Fig. 1. PRISMA flow diagram.

2.4. Effect size calculation and heterogeneity

For our first objective between-group effect sizes were calculated using Hedges's *g*. As in case of Cohen's *d* coefficient, a value of Hedge's *g* between 0.20 and 0.50 indicates a small effect, one between 0.50 and 0.80 indicates a medium effect, while a value larger than 0.80 indicates a large effect size (Cohen, 1988). In order to compute effect sizes, we used mean scores, standard deviations, and sample size. When there were studies that did not provide means and standard deviations we calculated *g* values from exact *t*, *F*, and *p* values applying conversion formulas when necessary. Thus, we obtained estimates of the effect and not the true effect as would be derived from means and standard deviations. We computed an average effect size for each study and used the study as the unit of analysis. Positive effect sizes indicated the advantage of virtual reality-based measures while negative effect sizes indicated the advantage of classical and computerized measures. Effect sizes were computed using random effects model which assumes two sources of variance: one is within study error, and second, variation in true effects across studies (Borenstein, Hedges, Higgins, & Rothstein, 2009). To test for heterogeneity of the effect sizes, we considered two statistics: the homogeneity test *Q* and the I^2 index.

Next, we performed subgroup analysis for executive functions measures, memory measures, and other neurocognitive measures, using fixed effect model given that there were few studies in each category (Borenstein et al., 2009). Although applying a random-effect meta-analysis is more realistic, produces more generalizable results and is highly recommended since we expect between-studies variance due to a high heterogeneity across samples of populations, when dealing with a reduced number of studies the procedure is not recommended because the between-studies variance estimated is unreliable (Borenstein et al., 2009).

2.5. Publication bias

Publication bias was investigated using Duval and Tweedie's trim-and-fill procedure (Duval & Tweedie, 2000). Trim-an-fill procedure identifies studies with extreme effect sizes from one side of the funnel plot and re-computes the effect sizes taking into account hypothetical symmetrical counterparts of those extremes. This way offers an unbiased estimate of the effect size.

2.6. Software

The statistical analysis was conducted using Comprehensive Meta-Analysis software (version 2.2, Borenstein, Hedges, Higgins, & Rothstein, 2005).

3. Results

For the first objective, the average effect sizes were calculated from 13 studies (N = 419), two that used a between-subject design (Gamberini, 2000; Lo Priore, Castelnuovo, Liccione, & Liccione, 2003), and 11 that used a within-subject design. The resulted effect sizes for the between-subject design was adjusted using Olejnik and Algina (2000) technical specifications. Results showed significant differences between virtual reality measures and computerized or classical measures with a medium effect size in favor of classical or computerized measures ($g = -0.77$, 95% CI [−1.29, −0.26], $z = -2.95$; $p = .003$). There was also evidence of high heterogeneity ($Q_{(12)} = 138.08$, $p < .001$; $I^2 = 91.31\%$). The negative sign indicates that classical or computerized assessments yield better performance. We addressed the high heterogeneity in the results by performing moderation analysis. Table 1 offers a synthetic view of the studies' characteristics and the forest plot in Fig. 2 displays the effect size values and 95% CI.

Further on, we computed average effect sizes for each category: executive functions measures and memory measures. In case of other neurocognitive measures, only one study was included, so that we could not compute any mean effect size and decided to report the effect size.

We calculated mean effect sizes for executive functions measures considering data from nine studies (N = 353). Results pointed out a significant difference between virtual reality measures and computerized or classic tests with a medium mean effect size ($g = -0.72$, 95% CI [−0.86, −0.58], $z = -10.14$; $p < .001$) with high heterogeneity ($Q_{(8)} = 64.36$, $p < .001$; $I^2 = 87.57\%$). The negative result indicates that performance on classical or computerized measures is better than performance on virtual reality measures. Next, we computed an average effect size for memory measures on data reported in three studies (N = 46). Results indicated significant differences between classical or computer-based measurement instruments and virtual reality instruments, in favor of virtual reality-based instruments, with a large effect size ($g = 1.65$, 95% CI [1.07, 2.24], $z = 5.59$; $p < .001$). Moreover, considering the increased heterogeneity ($Q_{(2)} = 11.37$, $p = .003$; $I^2 = 82.42\%$) and the fact that there is a considerable difference between the effect sizes of the three studies, results should be interpreted with caution (see Fig. 2). Third, in case of other neurocognitive measures, only one study was available, with a small negative effect size favoring classical measures ($g = -0.18$).

Given the high variability and scarce studies on these outcomes, we decided to report mean effect sizes for each of the following categories without providing data on statistical significance. Effect sizes for the comparison of classical paper-and-pencil or computer-based measures with virtual reality-based measures for each cognitive process are presented in Table 2.

3.1. Moderation analysis

The overall effect size for between-group analysis for cognitive performance on classical or computerized tests and virtual reality measures displayed statistically significant heterogeneity. In order to identify and explain the observed variability in effect sizes, we performed meta-regression and subgroup analysis.

The first potential moderator was participants' age which significantly moderated the overall effect size, with a tendency to stronger effects in case of younger participants. The strength of the mean weighted effect size tends to increase as the age of the participants decreases. This result should be interpreted with caution given the value of β, which indicates that the practical significance of the effect is null (see Table 3). Next, gender did not significantly moderate the effect size.

Subgroup analysis identified the clinical status of the sample as a moderator. To be more specific, the magnitude of the overall effect size was influenced by the type of participants included in the sample: clinical or healthy controls. Both types of participants yielded moderate effect sizes, with the strongest effect in the case of healthy participants. Another subgroup analysis was performed to see whether the type of control measurement instrument moderated the mean weighted effect size. The type of assessment tools did not moderate the effect size. The task performance indicator was a significant moderator of the effect size for the comparison on cognitive performance between virtual reality measures and classical or computerized tests (see Table 4).

3.2. Publication bias

We used Duval and Tweedie (2000) trim-and-fill procedure in order to investigate the presence of publication bias that estimated that no studies are missing which could modify the results. Such a result indicates that our results are robust and not affected by publication bias.

4. Discussion

The present meta-analysis investigated the task difficulty of virtual reality-based measures in comparison to classical paper-and-pencil or computerized cognitive measures. The present research dealt mainly with the cognitive performance measured either by virtual reality measures or analogue classical paper-and-pencil or computerized measures in order to examine the task difficulty hypothesis derived from the complexity of virtual reality measures. Overall, findings from this meta-analysis supported its' main purpose and provided evidence for the task complexity hypothesis of virtual reality-based measures.

Results point out significant differences between the two conditions with superior performance on classical and computerized psychometrics ($g = -0.77$, 95% CI [−1.29, −0.26]). These results are in line with our theoretical assumptions that virtual reality-based tests have an increased task difficulty compared to classical or computerized tests (Elkind et al., 2001; Gamberini, 2000). There are several explications for this pattern. First, is possible that the administration of neuropsychological tests via virtual reality triggers more cognitive resources. Second, virtual reality-based measure might be more demanding for the participants in comparison to classical and computerized measures because they replicate real world environments with stressors, distractors, and complex stimuli. As so, the examinee has to manipulate and process a larger amount of information, while completing the assessment tasks (Elkind et al., 2001; Rizzo et al., 2006). The contributions of all this factors which are specific to virtual reality environments could

Table 1
Characteristics of the studies included in the meta-analysis.

Author(s)	Mean age (years)	% Of male participants	N	Clinical status of the sample	Type of control measurement instrument	Type of control measurement instrument	Type of cognitive process assessed	Outcome measure	Type of VR platform	Effect size (Hedges's g)
Armstrong, Reger, Edwards, Rizzo, Courtney, and Parsons (2013)	28.78	93.90	49	Healthy	Classical paper-and-pencil, Computer-based	Time-based measures	Executive functions (attention)	VRST Color naming, complex interference, interference, word reading, D-KEFS Color naming, complex interference, interference, word reading, ANAM Color naming, interference, word reading	HMD	−2.52
Broeren et al. (2007)	54.37	50	8	Clinical (brain injury)	Classical paper-and-pencil	Error-based measures	Other neurocognitive measures (visuospatial neglect)	Star cancellation Visuospatial neglect, VR task visuospatial neglect	HMD	−0.18
Elkind et al. (2001)	29	75.02*	63	Healthy	Classical paper-and-pencil	Error-based measures	Executive functions	LFAM Conceptual responses, failure to maintain set, nonperseverative errors, perseverative errors, total error, trials to first category, WCST Conceptual responses, failure to maintain set, nonperseverative errors, perseverative errors, total error, trials to first category	HMD	−0.51
Gamberini (2000)	23.63	50	16	Healthy	Computer-based	Error-based measures	Memory (incidental memory)	VR environment Location task, recognition task, Desktop environment Location task, recognition task	HMD	−0.45
Nolin, Martin, and Bouchard (2009)	21.81*	75.02*	8	Clinical (brain injury)	Computer-based	Time-based measures, Error-based measures	Executive functions (attention)	VR Classroom Commissions, omissions, reaction time, VIGIL CPT Commissions, omissions, reaction time	HMD	−1.17
Lo Priore et al. (2003)	21.81*	75.02*	12	Healthy	Computer-based	Error-based measures	Memory (incidental memory)	V-STORE number of recalled elements presented in VR, V-	HMD	−0.81

(*continued on next page*)

Table 1 (*continued*)

Author(s)	Mean age (years)	% Of male participants	N	Clinical status of the sample	Type of control measurement instrument	Type of control measurement instrument	Type of cognitive process assessed	Outcome measure	Type of VR platform	Effect size (Hedges's g)
								STORE number of recalled elements presented in desktop environment		
Parsons and Courtney (2014)	25.58	75	50	Healthy	Classical paper-and-pencil	Time-based measures, Error-based measures	Executive functions (attention)	VR-PASAT Correct responses, response time, correct percent, PASAT-200 Correct responses, response time, correct percent	HMD	−0.44
Parsons et al. (2011)	21.81*	75.02*	20	Healthy	Classical paper-and-pencil, Computer-based	Error-based measures	Executive functions (attention)	VRST Correct percent, ANAM Correct percent, P&p Stroop Task Correct percent	HMD	−1.53
Parsons et al. (2013)	19.71	25	50	Healthy	Classical paper-and-pencil, Computer-based	Time-based measures, Error-based measures	Executive functions (attention)	VRST reaction time, number of correct responses on color-word and interference, D-KEFS reaction time, number of correct responses on color-word, interference and complex interference, ANAM reaction time, number of correct responses on color-word and interference	HMD	−4.65
Parsons et al. (2012)	21.81*	94	49	Healthy	Classical paper-and-pencil	Error-based measures	Executive functions (attention)	VR-PASAT Correct responses, PASAT-200 Correct responses	HMD	−0.81
Pollak, Shomaly, Weiss, Rizzo, and Gross-Tsur (2010)	13.70	59.25	27	Clinical (ADHD)	Computer-based	Time-based measures, Error-based measures	Executive functions (attention)	VR Classroom Commisions, omissions, reaction time, variability of reaction time, TOVA CPT Commisions, omissions, reaction time, variability of reaction time	HMD	−0.64
Pollak et al. (2009)	12.60	100	37	Healthy, Clinical (ADHD)	Computer-based	Time-based measures, Error-based measures	Executive functions (attention)	VR Classroom Commisions, omissions, reaction time, variability of reaction time, No VR Classroom Commisions,	HMD	−0.32

								omissions, reaction time, variability of reaction time	
Pugnetti, Mendozzi, Attree, Barbieri, Brooks, Cazzullo, Motta, and Rose (1998)	27.50	56.66	30	Healthy	Classical paper-and-pencil	Error-based measures	Memory (incidental memory)	VR Correct responses, Classic Correct responses	HMD	2.09

Note. Total $N = 301$; ANAM = Automated Neuropsychological Assessment Metrics-Fourth Edition (Reeves, Kane, Winter, & Goldstone, 1995); Desktop environment = nonimmersive desktop environment for object recognition task and object location task (Gamberini, 2000); D-KEFS = Delis-Kaplan Executive Function System (Delis, Kaplan, & Kramer, 2001); Classic correct responses = Classical paper-and-pencil for incidental memory assessment (Pugnetti et al., 1998); LFAM = Look for a Match (Elkind et al., 2001); No VR Classroom = Virtual Classroom (Rizzo et al., 2000) without immersion; PASAT-200 = Paced Auditory Serial Addition Test (Diehr, Heaton, Miller, & Grant, 1998); P&p Stroop Task = Paper-and-pencil Stroop Test (Stroop, 1935); Star cancellation = subtest in the Behavioral Inattention Test Battery (Halligan, Marshall, & Wade, 1989); TOVA CPT = Test of Variables of Attention (Greenberg & Waldmant, 1993); VRST = Virtual Reality Stroop Task (Parsons et al., 2011); V-STORE = immersive Virtual Reality-based tool (Lo Priore et al., 2003); VIGIL CPT = VIGIL Continous Performance Test (Cegalis, 1996); VR Classroom = Virtual Classroom (Rizzo et al., 2006, 2000); VR Correct responses = Virtual reality task for incidental memory assessment (Pugnetti et al., 1998); VR environment = Virtual Reality Environment for object recognition task and object location task; (Gamberini, 2000) VR-PASAT = Virtual Reality Paced Auditory Serial Addition Test (Parsons & Courtney, 2014); Virtual Reality VR task visuospatial neglect = Cancellation test developed in the Virtual reality environment (Broeren et al., 2007); WCST = Wisconsin Card Sorting Test (Grant & Berg, 1948); * = Mean age and mean of % of male participants were not provided in the studies and were substitute with the non-missing mean age and mean of % percentage of male participants of the studies included in the meta-analysis.

make virtual-reality-based tests more difficult for examinees and explain the results obtained in the current meta-analysis.

In order to investigate whether the pattern described above replicates among distinctive cognitive processes we performed additional analysis and compared the performance on executive functions, memory, and other neurocognitive measures. Between-group analysis revealed mixed results. For executive functions measures, results showed significant differences between classical and computerized measures and virtual reality measures ($g = -0.72$, 95% CI [−0.86, −0.58]). Again, cognitive performance assessed by classical or computerized measures was better than performance assessed via virtual reality measures. In case of memory measures, better performances were obtained with virtual reality measures, which points out that virtual reality tests seem to be easier than classical or computerized measures in case of memory assessment ($g = 1.65$, 95% CI [1.07, 2.24], $z = 5.59$; $p < .001$). Nevertheless, only three studies were available for analysis, so the results and the inferences made are not fully reliable. Moreover, one of the studies compared classical paper-and-pencil assessment with virtual reality (Pugnetti et al., 1998) and obtained a large effect size, in favor of virtual reality measures. The other two studies (Gamberini, 2000; Lo Priore et al., 2003) revealed small and large effect sizes, in favor of computerized measures. Finally, for other neurocognitive measures only one study was available showing a small effect size, in favor of classical measures ($g = -0.18$).

Because of the theoretical and practical importance we computed distinct comparisons of cognitive performance on virtual reality based measures with both classical paper-and-pencils measures and computerized measures, Results pointed out larger effect sizes for the comparison between computerized measures and virtual reality measures, with the superiority of computerized measures ($g = -0.86$). For the comparison between classical paper-and-pencil measures and virtual reality measures, results indicate a moderate mean effect size, in favor of classical measures ($g = -0.57$). Overall, virtual reality-based measures have an increased task difficulty that requires additional cognitive resources compared to both classical and computerized measures. However, for executive functions there is a larger effect size for the comparison between virtual reality measures and computerized measures compared to the effect size between virtual reality measures and classical measures. Nevertheless, the direction of effect size points out to an increased task difficulty of virtual reality based measures. In case of memory assessment, we obtained mixed results. Overall, it seems that for memory assessment tasks embedded in virtual reality are easier, although taken into consideration the type of assessment instrument, results suggest that compared to classical paper-and-pencil measures virtual reality assessment has a low level of task difficulty and complexity, while for computerized measures the pattern is reversed, and virtual reality tests are more complex and difficult (see Table 4). It might be possible that memory tasks embedded in virtual reality offer more cues for retention because they present more realistic mental images which resemble everyday situations. However, delivered via HMDs the virtual world becomes more complex, as well as the amount and complexity of information to be processed. Consequently, performance tends to decrease compared to computerized assessment. Nevertheless, these interpretations are based on only three studies so the reliability of inferences made is limited.

4.1. Moderator effects

All of the main effects in the meta-analysis revealed heterogeneity and as a consequence we focused on our second objective on moderation analysis.

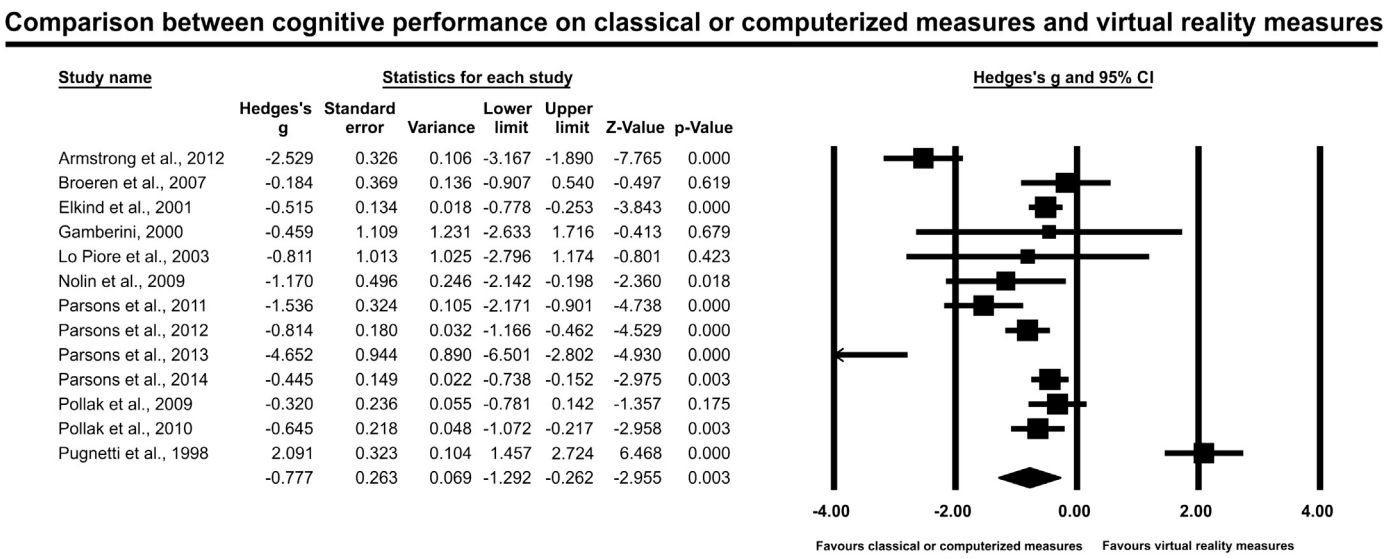

Fig. 2. Comparison between cognitive performance on classical or computerized measures and virtual reality measures.

Table 2
Mean effect sizes (Hedges's g) for executive functions, memory and other neurocognitive measures depending on type of control measurement instrument (classical paper-and-pencil measures and computer-based measures versus virtual reality measures).

Categories based on the following cognitive process assessed	Hedges's *g* for classical paper-and-pencil measures versus virtual reality measures	Hedges's *g* for computer-based measure versus virtual reality measures	Hedges's *g* for classical or computerized measures versus virtual reality measures
Executive functions	−0.76 ($K = 6$)	−0.86 ($K = 6$)	−0.72 ($K = 9$)
Memory	2.09 ($K = 1$)	−0.47 ($K = 2$)	1.65 ($K = 3$)
Other neurocognitive measures	−0.18 ($K = 1$)	($K = 0$)	−0.18 ($K = 1$)
Total	−0.57 ($K = 8$)	−0.86 ($K = 8$)	−0.77 ($K = 13$)

Note. K = number of studies included in the analysis.

Table 3
Meta-regression analysis with numeric variables for performance on cognitive measures.

Outcome	Moderator	*K*	β	Standard error	95% CI	z	p	Q model	p
Performance on virtual reality cognitive measures	Age	13	−0.014	0.00	[−0.02; 0.00]	−3.13	0.001	9.81	0.001
	Gender	13	−0.001	0.00	[−0.00; −0.00]	−0.58	0.000	26.44	0.555

Note. K = number of studies included in the analysis; 95% CI = 95% confidence interval around the weighted mean effect size.

Table 4
Moderation analysis with categorical variables for performance on cognitive measures.

Outcome	Moderator	*K*	*g*	*p*	Q w	*p*	*95%* CI	Q b	*p*
Performance on cognitive measures	Healthy/Clinic	9	−0.66	0.000	496.28	0.000	[−0.74; −0.59]	7.31	0.007
		4	−0.48	0.000	46.17	0.000	[−0.59;−0.36]		
				0.000					
	Classic/	8	−0.59	0.000	313.08	0.000	[−0.67; −0.51]	0.98	0.321
	Computer	7	−0.65	0.000	235.70	0.000	[−0.75; −0.55]		
	Time-based measures/	6	−0.84	0.000	332.38	0.000	[−0.95; −0.73]	26.08	0.000
	Error-based measures	11	−0.50	0.000	191.29	0.000	[−0.58; −0.43]		

Note. K = number of studies included in the analysis; g = Hedge's g; 95% CI = 95% confidence interval around the weighted mean effect size.

The first significant moderator was participants' age. The more the age of the participants increased, the effect size decreased so that the difference in cognitive performance between virtual reality measures and classical or computerized measures have been reduced. This could mean that the cognitive decline associated with an increase in age makes the performance obtained on both types of assessment instruments more similar. The second significant moderator was the clinical status of the sample. As anticipated, results pointed out stronger effects for healthy participants. Larger effect sizes in case of healthy controls in comparison to clinical samples can be explained by the fact that cognitive impairment associated with clinical condition will shorten the effects accounted for task difficulty. Next, task performance indicator moderates the overall effect size, which indicates time-based measures account for larger differences between the virtual reality measures and classical or computerized tests than error-based measures. Time-

based measures may be more sensitive to measurement procedure or require additional cognitive resources.

5. Limitations and conclusions

The findings presented in this meta-analysis have several shortcomings. The first limitation refers to the small number of studies that were included in the analysis which may weaken the statistical power. This drawback also reflected in the moderation analysis as there were no sufficient studies to test for all the potential theoretical moderators. In some cases, subgroup analysis was performed with a small sample size of effects from primary studies which may affect the robustness and reliability of analysis. Furthermore, when comparing cognitive performance on virtual reality measures with classical and computerized assessment tests, there were insufficient studies to perform different comparisons for each of the pairs. We were able to provide only a mean effect sizes for each comparison without any data on statistical significance or heterogeneity.

Future research should focus more on predictive validity of virtual reality-based measures in relationship to real-life performance or other objective criteria and to investigate the equivalence or superiority in task performance of either measure. Also, studies might consider providing norms and reliability analysis for virtual reality-based measures, as well as more reliable indexes of classification accuracy, such as sensitivity, specificity, positive predictive power, and negative predictive power.

Overall, results from the current meta-analysis point out that cognitive performance obtained in virtual reality is poorer than the one in classical or computerized assessment which might suggest that tasks embedded in virtual reality have an increased level of complexity and difficulty and require additional cognitive resources.

Author disclosure statement

No competing financial interests exist.

Acknowledgment

“This work was possible due to the financial support of the Sectorial Operational Program for Human Resources Development 2007-2013, co-financed by the European Social Fund, under the project number POSDRU/159/1.5/S/132400 with the title Young successful researchers – professional development in an international and interdisciplinary environment.”

References[1]

Adams, R., Finn, P., Moes, E., Flannery, K., & Rizzo, A. A. (2009). Distractibility in attention/deficit/hyperactivity disorder (ADHD): the virtual reality classroom. *Child Neuropsychology, 15*(2), 120–135. http://dx.doi.org/10.1080/09297040802169077.

Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: a meta-analytic review. *Neuropsychology Review, 16*(1), 17–42. http://dx.doi.org/10.1007/s11065-006-9002-x.

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

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist, 26*(2), 177–196. http://dx.doi.org/10.1093/arclin/acs027.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2005). *Comprehensive meta-analysis (Version 2)* [Computer Software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, England: Wiley.

*Broeren, J., Samuelsson, H., Stibrant-Sunnerhagen, K., Blomstrand, C., & Rydmark, M. (2007). Neglect assessment as an application of virtual reality. *Acta Neurologica Scandinavica, 116*(3), 157–163. http://dx.doi.org/10.1111/j.1600-0404.2007.00821.x.

Cegalis, J. A. (1996). *VIGIL continuous performance test. User's guide*. San Antonio, TX: Psychological Corporation.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Diehr, M. C., Heaton, R. K., Miller, W., & Grant, I. (1998). The paced auditory serial addition task (PASAT): norms for age, education, and ethnicity. *Assessment, 5*(4), 375–387. http://dx.doi.org/10.1177/107319119800500407.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x.

Elkind, J. S. (1998). Use of virtual reality to diagnose and habilitate people with neurological dysfunctions. *CyberPsychology & Behavior, 1*(3), 263–273. http://dx.doi.org/10.1089/cpb.1998.1.263.

*Elkind, J. S., Rubin, E., Rosenthal, S., Skoff, B., & Prather, P. (2001). A simulated reality scenario compared with the computerized Wisconsin Card sorting test: an analysis of preliminary results. *CyberPsychology & Behavior, 4*(4), 489–496. http://dx.doi.org/10.1089/109493101750527042.

*Gamberini, L. (2000). Virtual reality as a new research tool for the study of human memory. *CyberPsychology & Behavior, 3*(3), 337–342. http://dx.doi.org/10.1089/10949310050078779.

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

Greenberg, L. M., & Waldmant, I. D. (1993). Developmental normative data on the test of variables of attention (TOVA™). *Journal of Child Psychology and Psychiatry, 34*(6), 1019–1030. http://dx.doi.org/10.1111/j.1469-7610.1993.tb01105.x.

Halligan, P. W., Marshall, J. C., & Wade, D. T. (1989). Visuospatial neglect: underlying factors and test sensitivity. *The Lancet, 334*(8668), 908–911. http://dx.doi.org/10.1016/S0140-6736(89)91561-4.

Henry, M., Joyal, C. C., & Nolin, P. (2012). Development and initial assessment of a new paradigm for assessing cognitive and motor inhibition: the bimodal virtual-reality Stroop. *Journal of Neuroscience Methods, 210*(2), 125–131. http://dx.doi.org/10.1016/j.jneumeth.2012.07.025.

Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., & Ziefle, M. (2011a). Investigating paper vs. screen in real-life hospital workflows: performance contradicts perceived superiority of paper in the user experience. *International Journal of Human-Computer Studies, 69*(9), 563–570. http://dx.doi.org/10.1016/j.ijhcs.2011.05.002.

Holzinger, A., Searle, G., & Wernbacher, M. (2011b). The effect of previous exposure to technology on acceptance and its importance in usability and accessibility engineering. *Universal Access in the Information Society, 10*(3), 245–260. http://dx.doi.org/10.1007/s10209-010-0212-x.

Ku, J., Cho, W., Kim, J. J., Peled, A., Wiederhold, B. K., Wiederhold, M. D., et al. (2003). A virtual environment for investigating schizophrenic patients' characteristics: assessment of cognitive and navigation ability. *CyberPsychology & Behavior, 6*(4), 397–404. http://dx.doi.org/10.1089/109493103322278781.

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. http://dx.doi.org/10.1016/j.jneumeth.2013.07.005.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York, NY: Oxford University Press.

*Lo Priore, C., Castelnuovo, G., Liccione, D., & Liccione, D. (2003). The creation of V-STORE. In B. K. Wiederhold, & G. Riva (Eds.), *Annual review of cybertherapy and telemedicine* (Vol. 1, pp. 29–36). Amsterdam, NL: IOS Press.

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

Myers, R. L., & Bierig, T. A. (2000). Virtual reality and left hemineglect: a technology for assessment and therapy. *CyberPsychology & Behavior, 3*(3), 465–468. http://dx.doi.org/10.1089/10949310050078922.

*Nolin, P., Martin, C., & Bouchard, S. (2009). Assessment of inhibition deficits with the virtual classroom in children with traumatic brain injury: a pilot-study. *Annual Review of CyberTherapy and Telemedicine, 144*, 240–242. http://dx.doi.org/10.3233/978-1-60750-017-9-240.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*(3), 241–286. http://dx.doi.org/10.1006/ceps.2000.1040.

Parsons, T. D. (2012). Virtual simulations and the second life metaverse: paradigm shift in neuropsychological assessment. In N. Zagalo, L. Morgado, & A. Boa-Ventura (Eds.), *Virtual worlds and metaverse platforms: New communication and identity paradigms* (pp. 234–250). Hershey, PA: Information Science Reference. http://dx.doi.org/10.4018/978-1-60960-854-5.ch016.

[1] References marked with an asterisk indicate studies included in the meta-analysis.

*Parsons, T. D., & Courtney, C. G. (2014). An initial validation of the virtual reality paced auditory serial addition test in a college sample. *Journal of Neuroscience Methods, 222*, 15–23. http://dx.doi.org/10.1016/j.jneumeth.2013.10.006.

*Parsons, T. D., Courtney, C. G., Arizmendi, B. J., & Dawson, M. E. (2011). Virtual reality stroop task for neurocognitive assessment. In J. D. Westwood (Ed.), *Medicine meets virtual reality* (pp. 433–439). Amsterdam, NL: IOS Press.

*Parsons, T. D., Courtney, C. G., & Dawson, M. E. (2013). Virtual reality Stroop task for assessment of supervisory attentional processing. *Journal of Clinical and Experimental Neuropsychology, 35*(8), 812–826. http://dx.doi.org/10.1080/13803395.2013.824556.

*Parsons, T. D., Courtney, C. G., Rizzo, A. A., Armstrong, C., Edwards, J., & Reger, G. (2012). Virtual reality paced serial assessment test for neuropsychological assessment of a military cohort. In J. D. Westwood (Ed.), *Medicine meets virtual reality* (pp. 331–337). Amsterdam, NL: IOS Press.

Parsons, T. D., Larson, P., Kratz, K., Thiebaux, M., Bluestein, B., Buckwalter, J. G., et al. (2004). Sex differences in mental rotation and spatial rotation in a virtual environment. *Neuropsychologia, 42*(4), 555–562. http://dx.doi.org/10.1016/j.neuropsychologia.2003.08.014.

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

Podell, K., DeFina, P. A., Barrett, P., McCullen, A., & Goldberg, E. (2003). Assessment of neuropsychological functioning. In J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology* (pp. 443–466). New Jersey, NJ: Wiley.

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

*Pollak, Y., Weiss, P. L., Rizzo, A. A., Weizer, M., Shriki, L., Shalev, R. S., et al. (2009). The utility of a continuous performance test embedded in virtual reality in measuring ADHD-related deficits. *Journal of Developmental & Behavioral Pediatrics, 30*(1), 2–6. http://dx.doi.org/10.1097/DBP.0b013e3181969b22.

*Pugnetti, L., Mendozzi, L., Attree, E. A., Barbieri, E., Brooks, B. M., & Cazzullo, C. L. (1998). Probing memory and executive functions with virtual reality: past and present studies. *CyberPsychology & Behavior, 1*(2), 151–162. http://dx.doi.org/10.1089/cpb.1998.1.151.

Rand, D., Basha-Abu Rukan, S., Weiss, P. L., & Katz, N. (2009). Validation of the virtual MET as an assessment tool for executive functions. *Neuropsychological Rehabilitation: An International Journal of Head Trauma Rehabilitation, 19*(4), 583–602. http://dx.doi.org/10.1080/09602010802469074.

Rand, D., Kizony, R., Feintuch, U., Katz, N., Josman, N., Rizzo, A., et al. (2005). Comparison of two VR platforms for rehabilitation: Video capture versus HMD. *Presence, 14*(2), 147–160. http://dx.doi.org/10.1162/1054746053967012.

Reeves, D., Kane, R. L., Winter, K. P., & Goldstone, A. (1995). *Automated neuropsychological assessment metrics (ANAM V3. 11): Clinical and neurotoxicology subsets*. San Diego, CA: National Cognitive Recovery Foundation (Report No. NCRF-SR-95–01).

Rheingold, H. (1991). *Virtual reality*. New York, NY: Summit.

Riva, G. (1998). Virtual reality as assessment tool in psychology. In G. Riva (Ed.), *Virtual reality in neuro-psycho-physiology* (pp. 71–79). Amsterdam, NL: IOS Press.

Rizzo, A. A., Bowerly, T., Buckwalter, J. G., Klimchuk, D., Mitura, R., & Parsons, T. D. (2006). A virtual reality scenario for all seasons: the virtual classroom. *CNS Spectrums, 11*(1), 35–44. http://dx.doi.org/10.1017/S1092852900024196.

Rizzo, A. A., Buckwalter, J. G., Bowerly, T., Van Der Zaag, C., Humphrey, L., Neumann, U., et al. (2000). The virtual classroom: a virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior, 3*(3), 483–499. http://dx.doi.org/10.1089/10949310050078940.

Rizzo, A. A., Buckwalter, J. G., Neumann, U., Chua, C., Van Rooyen, A., Larson, P., et al. (1999). Virtual environments for targeting cognitive processes: an overview of projects at the University of Southern California. *CyberPsychology & Behavior, 2*(2), 89–100. http://dx.doi.org/10.1089/cpb.1999.2.89.

Rizzo, A. A., Schultheis, M., Kerns, K. A., & Mateer, C. (2004). Analysis of assets for virtual reality applications in neuropsychology. *Neuropsychological Rehabilitation, 14*(1–2), 207–239. http://dx.doi.org/10.1080/09602010343000183.

Schultheis, M. T., Himelstein, J., & Rizzo, A. A. (2002). Virtual reality and neuropsychology: upgrading the current tools. *Journal of Head Trauma Rehabilitation, 17*(5), 378–394. http://dx.doi.org/10.1097/00001199-200210000-00002.

Sorita, E., N'Kaoua, B., Larrue, F., Criquillon, J., Simion, A., Sauzéon, H., et al. (2013). Do patients with traumatic brain injury learn a route in the same way in real and virtual environments? *Disability and Rehabilitation, 35*(16), 1371–1379. http://dx.doi.org/10.3109/09638288.2012.738761.

Stickel, C., Ebner, M., & Holzinger, A. (2010). The XAOS metric – understanding visual complexity as measure of usability. In G. Leitner, M. Hitz, & A. Holzinger (Eds.), *HCI in work and learning, life and leisure, lecture notes in computer science* (Vol. 6389, pp. 278–290). Berlin, Heidelberg: Springer.

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

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey, NJ: Wiley.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250–270. http://dx.doi.org/10.1037/0033-2909.117.2.250.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology* (pp. 43–66). New Jersey, NJ: Wiley.