JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY, 2013
Vol. 35, No. 2, 113–123, http://dx.doi.org/10.1080/13803395.2012.740002

Validity of the Virtual Reality Stroop Task (VRST) in active duty military

Christina M. Armstrong[1], Greg M. Reger[1], Joseph Edwards[1], Albert A. Rizzo[2], Christopher G. Courtney[2], and Thomas D. Parsons[2] thomas.parsons@unt.edu

[1]National Center for Telehealth and Technology (T2) Defense Centers of Excellence (DCoE) for Psychological Health & Traumatic Brain Injury, Joint Base Lewis-McChord, Tacoma, WA, USA
[2]Institute for Creative Technologies, University of Southern California, Los Angeles, CA, USA

Virtual environments provide the ability to systematically deliver test stimuli in simulated contexts relevant to real world behavior. The current study evaluated the validity of the Virtual Reality Stroop Task (VRST), which presents test stimuli during a virtual reality military convoy with simulated combat threats. Active duty Army personnel ($N = 49$) took the VRST, a customized version of the Automated Neuropsychological Assessment Metrics (ANAM)–Fourth Edition TBI Battery (2007) that included the addition of the ANAM Stroop and Tower tests, and traditional neuropsychological measures, including the Delis-Kaplan Executive Function System version of the Color–Word Interference Test. Preliminary convergent and discriminant validity was established, and performance on the VRST was significantly associated with computerized and traditional tests of attention and executive functioning. Valid virtual reality cognitive assessments open new lines of inquiry into the impact of environmental stimuli on performance and offer promise for the future of neuropsychological assessments used with military personnel.

Keywords: Stroop test; Executive functioning; Virtual reality; Validity, Active duty military.

INTRODUCTION

Traumatic brain injury (TBI) is a significant health issue that affects many service members and veterans. Between 2000 and 16 May 2011, a total of 212,742 service members sustained a traumatic brain injury in support of Operation Iraqi Freedom (OIF), Operation Enduring Freedom (OEF), and Operation New Dawn (OND) (Armed Forces Health Surveillance Center, 2011). In a recent study of soldiers deployed to Iraq, clinician-confirmed TBI history (primarily mild traumatic brain injury, or mTBI) was identified in more than one of every five (22.8%) Army soldiers from a Brigade Combat Team (Terrio et al., 2009), while these rates have varied between 8% and 23% (Vasterling et al., 2006).

Although the majority of individuals with mTBI do not have measurable long-term neurocognitive deficits and have a full recovery, as many as 7–33% of mTBI patients develop one or more persistent somatic, cognitive, emotional or behavioral post-concussive symptoms that may be related to traumatic brain alteration (Alexander, 1995; Belanger, Curtiss, Demery, Lebowitz, & Vanderploeg, 2005; Deb, Lyons, Koutzoukis, Ali, & McCarthy, 1999; Hofman et al., 2001; Hofman, Verhey, Wilmink, Rozendaal, & Jolles, 2002; Hoge et al., 2008; Ryan

We would like to thank Lisa Thomas and Emily Fantelli for their support in this project. We would also like to acknowledge Telemedicine and Advanced Technology Research Center (TATRC), which funded a portion of this research. Thomas D. Parsons changed positions and is now at Clinical Neuropsychology and Simulation Lab, University of North Texas, Denton, TX, USA. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or reflecting the views of the Department of the Army or the Department of Defense.

Address correspondence to Christina M. Armstrong, National Center for Telehealth and Technology (T2) Defense Centers of Excellence (DCoE) for Psychological Health & Traumatic Brain Injury, Old Madigan Army Medical Center Bldg. 9933A Tacoma, WA 98431 (E-mail: drchristinaarmstrong@gmail.com).

& Warden, 2003). In some instances, subtle long-term effects on attention and working memory might persist for months or even years after the original injury (Bohnen & Jolles, 1992; Ponsford et al., 2000; Vanderploeg, Curtiss, & Belanger, 2005). Neuropsychological dysfunction that persists following mTBI may be best identified using assessments measuring attention (Binder, 1997), processing speed, working memory, memory, and executive functioning (Frencham, Fox, & Mayberry, 2005).

Military neuropsychologists use a number of paper-and-pencil neuropsychological measures to gauge cognitive deficits related to traumatic brain injury, including the Stroop Color–Word Interference test used to measure attention processes and executive functioning. Initially developed several decades ago (Stroop, 1935), the Stroop Color–Word Interference test has since been cited in hundreds of studies and is one of the most widely used neuropsychological assessments (McLeod, 1992). Many different versions of the Stroop test have been developed over the last seven decades, including auditory, picture–word, sorting, and matching versions (see McLeod 1991 for a review).

While traditional procedures for neuropsychological assessment of a soldier's executive functioning are still widely used, technological advances in other clinical neurosciences (e.g., the development of neuroimaging) have changed the neuropsychologist's role from differential diagnosis of brain pathology to that of also making predictions about the impact of a given person's neurocognitive abilities and disabilities on everyday functioning (Sbordone, 1996). Standard neurocognitive batteries tend to examine isolated components of neuropsychological ability, and thus may not adequately predict overall functioning that relate to difficult questions such as return to service following an injury (Parsons, Rizzo, & Buckwalter, 2004; Parsons, 2011).

Despite the many versions of the Stroop test and other neuropsychological assessment instruments that have been developed over the years, the vast majority of current neuropsychological assessment procedures represent a technology that has not changed since the first scales developed in the early 1900s (e.g., Binet and Simon's first scale in 1905 and Wechsler's in 1939). There are limitations in traditional methods of neuropsychological assessment such as variations in stimuli presentation and scoring (speed and accuracy) and the lack of ecological validity. Computerized versions of assessments provide the benefit of delivering stimuli systematically and the ability to track speed and accuracy with precision. Thus, validated computerized versions of traditional paper-and-pencil neuropsychological tests are in high demand. However, not even computerized versions of neuropsychological assessments show how cognitive functioning may change in stressful situations, such as during military combat. Neuropsychological assessments are traditionally administered in optimum environments and aim to obtain the best possible performance of the individual. However, it is also important to know how adaptive decision-making in stressful situations occurs, which may tell us about operational performance. Given the increasing prevalence of blast injuries to the head within the military, and the fact that many brain injuries may have no external marker of injury, researching innovative cognitive assessment methods in the assessment of cognitive functioning and detecting brain injury are a pressing need.

Virtual reality (VR) is an advanced computer interface that can allow a soldier to become immersed within a computer-generated simulation. Potential VR use in the assessment and training of cognitive processes in soldiers is becoming recognized as the need for assessments that are specifically designed for military service members increases and as the VR technology advances. Since virtual environments (VEs) allow for precise presentation and control of dynamic perceptual stimuli (visual, auditory, olfactory, ambulatory, and haptic conditions), they can provide assessments that combine the control and rigor of laboratory measures with a task similarity that better approximates real life situations.

VR applications that focus on component cognitive processes, including attention processes (Parsons, Cosand, Courtney, Iyer, & Rizzo, 2009; Parsons, Rizzo, Bamattre, & Brennan, 2007), spatial abilities (Parsons et al., 2004), memory (Parsons & Rizzo, 2008), and executive functions (Elkind, Rubin, Rosenthal, Skoff, & Prather, 2001), have been developed and tested in an effort to increase the ecological validity of neurocognitive batteries that include assessment to support differential diagnosis and treatment planning. Within a VE, it is possible to systematically present cognitive tasks to service members that target military relevant neuropsychological performance beyond what is currently available using traditional methods.

The Virtual Reality Cognitive Performance Test (VRCPAT) uses a VR system to administer a battery of cognitive tests. The Virtual Reality Stroop Task (VRST) is one of the assessments included in the VRCPAT battery and is a version of the traditional Stroop test (Stroop, 1935) that is set in a military-themed VE (i.e., a simulated Humvee

convoy). Preliminary psychometric properties of the VRST have been established in an adult civilian population (Parsons, Courtney, Arizmendi, & Dawson, 2011), but there is no research exploring its validity in a military population. In light of the growing need among neuropsychologists to develop and refine strategies for the development of improved cognitive assessments for use in military populations, the current study aims to establish the preliminary validity of the VRST in an active duty military sample.

It was hypothesized that (a) VRST would significantly correlate with established neuropsychological measures of attention and executive functioning; (b) that patterns of performance for word reading, color naming, and interference tests of the VRST would be similar to traditional Stroop tests; (c) that VRST would not significantly correlate with neuropsychological assessments assessing dissimilar cognitive domains (i.e., memory); (d) that VRST response times would be longer than ANAM and D-KEFS response times, since tasks performed in a virtual environment are thought to require additional cognitive resources; (e) VRST scores would be more highly correlated with the ANAM Stroop test than the D-KEFS Stroop test due to the similarities in the method of computerized administration.

METHOD

Participants

Active duty soldiers were recruited from a large military installation in the continental United States. Inclusion criteria included English-speaking active duty soldiers between the ages of 18 and 64, with at least a high school education or GED equivalent. Potential participants were excluded if they (a) screened positive for PTSD on the PTSD Checklist–Military version (PCL–M; total score >50), (b) had a previous head injury with loss of consciousness greater than 15 min since beginning active duty, (c) had a psychiatric or neurologic disorder known to adversely impact cognition, including psychotic disorders, ADHD, bipolar disorder, or organic brain disorders, (d) had a history of migraines cued by auditory or visual stimuli, (e) had a history of seizures, (f) had a history of a physical condition that interferes with the proper use of the virtual reality head-mounted display or its peripherals, (g) had a strong propensity for motion sickness, (h) had an inability to perform Stroop tasks due to color-blindness, (i) were female soldiers who were pregnant or breast-feeding.

Information regarding the current study was presented to active duty soldiers during a briefing and 149 expressed an initial interest in the study. Of the 149 soldiers who expressed an initial interest in the study, 56 were able to be contacted and attended a consenting appointment. All 56 consented to study procedures. Six of the soldiers who had consented were excluded for the following reasons: (a) history of migraines ($n = 2$), (b) history of significant motion sickness ($n = 1$), (c) diagnosis of bipolar disorder ($n = 1$), (d) color blindness ($n = 1$), and (e) diagnosis of ADHD, as well as being color-blind ($n = 1$). Of the 50 soldiers who were enrolled in the study, 3 experienced simulator sickness while conducting the virtual reality portion of study procedures, and one terminated participation. Two soldiers felt nauseous during the virtual reality portion of study procedures, and one soldier experienced headaches, both symptoms of simulator sickness. The soldier who terminated participation experienced simulator sickness symptoms quickly after beginning study procedures; thus no useable data were available. However, the majority of the data were collected for the other two participants, and they are included in the overall analyses. A final sample size included data for 49 active duty soldiers.

Demographic information is included in Table 1. The final study sample had a mean age of 28.78 years ($SD = 2.23$); they were mostly married (57.1%), lower-enlisted (61.2%) males (93.9%) with some college education (57.1%). The soldiers endorsed the following racial groups: (a) Caucasian (51%), (b) African-American (28.6%), Other (14.3%), and Pacific Islander (4.1%); 18% of the soldiers reported their ethnicity as Hispanic. The mean number of deployments for soldiers in the current study was 1.12 ($SD = 1.64$). Almost half of the soldiers ($n = 23$; 46.9%) had no previous deployments, nearly a quarter ($n = 12$, 24.5%) had one previous deployment, 9 (18.4%) had two previous deployments, and 2 (4.1%) had three previous deployments. Three soldiers (6.0%) had more than three previous deployments. The demographics of the current sample are comparable to those of the overall Army population (Department of Defense, 2009).

Measures

The battery of neurocognitive tests included the Delis-Kaplan Executive Function System (D-KEFS) version of the Color–Word Interference Test (Delis, Kramer, & Kaplan, 2001), the Paced Auditory Serial Addition Test (PASAT; Gronwall,

TABLE 1
Demographics

Characteristic	*Mean*	*SD*	*N*	*%*
Age	28.78	2.23		
# Years in Army	6.89	6.44		
# Deployments	1.12	1.64		
Sex (male)			46	93.9
Handedness				
Right			48	98.0
Left			0	0.0
Ambidextrous			1	2.0
Race				
Caucasian			25	51.0
African American			14	28.6
Pacific Islander			2	4.1
Other			7	14.3
Ethnicity (not Hispanic)			40	81.6
Marital status				
Single/never Married			12	24.5
Married			26	57.1
Divorced			6	12.2
Widowed			0	0.0
Separated			5	10.2
Education				
GED/HS diploma			13	26.5
Some college			28	57.1
AA/Technical degree			3	6.1
Bachelor's degree			2	4.1
Master's degree or higher			3	6.1
Rank				
E1-E4			30	61.2
E5-E9			18	36.7
O1-O5			1	2.1
O6-O10			0	0.0

Note. $N = 49$. E = Enlisted rank; O = Officer rank.

1977), the Automated Neuropsychological Assessment Metrics–Fourth Edition, TBI Battery (ANAM 4 TBI), and two virtual reality assessments from the Virtual Reality Cognitive Performance Assessment Test (VRCPAT), including the Virtual Reality Stroop Task (VRST) and the Virtual Reality City Memory PASAT.

Delis-Kaplan Executive Function System (D-KEFS) Color–Word Interference Test (Delis et al., 2001)

The D-KEFS Color-Word Interference Test is a version of the classic Stroop Color–Word task and has been shown to be a valid and highly reliable measure of executive attentional control (MacLeod, 1991; Stroop, 1935). The D-KEFS Color–Word Interference Test is scored by assessing how long a participant takes to complete each of four conditions, with 50 items per condition. The four conditions include color naming, word reading, inhibition, and inhibition/switching. In order for the D-KEFS Color–Word Interference Test scores to be comparable to the other versions of the Stroop test (ANAM and VRST) in the current study, total time to complete each condition was divided by the number of stimuli per condition (50), then multiplied by 1,000 in order to obtain a mean response time in milliseconds.

Paced Auditory Serial Addition Test (PASAT; Diehr, Heaton, Miller & Grant, 1998; Gronwall, 1977)

The PASAT requires the participant to listen to numbers and add together in consecutive fashion the last two digits heard. The 200-item form of the PASAT assesses basic memory, attention, and processing abilities with demonstrated validity and reliability in adult populations (Crawford, Obonsawin, & Allan, 1998; Diehr, Cherner, Wolfson, Miller, & Grant, 2003).

Automated Neuropsychological Assessment Metrics–Fourth Edition (ANAM 4)

The ANAM 4 is a computerized neuropsychological assessment battery. For the current study, a customized ANAM battery that included the ANAM TBI battery of tests with the addition of the Stroop Color–Word Interference Test and Tower Tests was used. Tests included in the current battery were: Reaction Time, Code Substitution, Procedural Reaction Time, Mathematical Processing, Matching to Sample, and Code Substitution–delayed, as well as the Stroop Color–Word Interference Test and the Tower Test. The ANAM has been well established as a valid and reliable test battery for use in military populations (Jones et al., 2008; Vincent et al., 2008).

Virtual Reality Stroop Task (VRST; Parsons, et al., 2011)

The VRST is a part of the VRCPAT battery and involves the presentation of Stroop stimuli while immersed in a virtual environment. Stimuli are superimposed on a virtual windshield of a military Humvee, while the Humvee automatically drives down a desert road in Iraq. Four conditions are presented, including word reading, color naming, interference, and a complex interference task Stroop stimuli are continually presented in a fixed central location on the windshield. During the complex interference condition the Stroop stimuli are presented randomly throughout the windshield.

Stroop stimuli presentation progresses based upon the participant's response of either "red," green," or "blue," which are captured via a color-coded keypad. Response time and accuracy are recorded. Preliminary validation of the VRST as an assessment of neurocognitive functioning has been conducted in a non-military population (Parsons, et al., 2011).

Virtual Reality City Memory PASAT (Parsons et al., 2012; Parsons, Silva, Pair, & Rizzo, 2008)

The VR City Memory PASAT is part of the VRCPAT battery and requires the user to navigate through a simulated Iraqi city environment. The stimuli are primarily language-based (i.e., blue vehicle with bullet holes in windshield, intact barrel with "U.S. Army" stenciled on it, olive shipping container with numbers on it, etc.). After the first phase, the participant is immersed in the virtual environment and asked to perform the PASAT test while walking to different "zones." Upon arrival at each zone, participants are asked to accurately identify and take virtual pictures of items in each zone that they recall from the acquisition phase learning trials (while ignoring distracter/nontarget items). The PASAT stimuli are first presented at the rate of one item per 3 s, and then presented at the rate of one item per 2 s. Following the immersion in the virtual environment, the participant is then asked to recall as many items from the original list as possible.

Procedures

Upon local Institutional Review Board approval, study procedures were conducted individually with participants in a private office on a large military installation. Following voluntary, written consent, participants were administered the Beck Depression Inventory–II (BDI–II; Beck, Steer, & Brown, 1996) and the Posttraumatic Stress Disorder Checklist–Military Version (PCL–M; Weathers, Huska, & Keane, 1991) for screening purposes. Participants not screening positive for PTSD completed a brief demographics questionnaire followed by the neurocognitive assessment battery. The battery of tests was counterbalanced using a Latin square design in order to minimize experimental bias due to sequential effects. Order of administration was randomized across participants. Consent procedures, screening measures, and administration of neurocognitive measures took approximately two hours to complete over one session.

Tests were administered by either a research psychologist with specialized education and training in neuropsychological assessment (first author) or research coordinators with extensive experience and training in the administration of the current test battery. All persons administering tests for the current study also had received specialized training in virtual reality testing procedures and were supervised by a licensed clinical psychologist (second author).

Data analysis

Several preliminary analyses were conducted prior to the examination of the main hypotheses. Demographic variables were examined in order to determine whether any relationships existed with the three versions of the Stroop (VRST, ANAM, and D-KEFS). Additionally, effects of test administration order and possible correlations with screening measures (BDI–II and PCL–M) were examined. Main hypotheses were analyzed using bivariate correlations. Magnitude of effect sizes and patterns of scores were examined in order to determine convergent and discriminant validity.

RESULTS

VRST, ANAM, and D-KEFS variables and demographics

There was no relationship between age and any of the D-KEFS conditions. Participants who were older responded more slowly on computerized tests of word reading and color naming. Specifically, age positively correlated with the average response times for the ANAM word reading condition (WR) ($r = .32$, $p < .05$), ANAM color naming condition (CN) ($r = .30$, $p < .05$), VRST WR ($r = .42$, $p < .01$), and VRST CN ($r = .37$, $p < .05$).

Differences by education emerged on ANAM WR, $F(1, 47) = 10.58$, $p < .01$ (see Table 2). Pairwise comparisons found that participants with a high school diploma or GED responded significantly faster than participants with some college or more education. Similarly, there were significant differences between education groups on ANAM CN, $F(1, 47) = 8.29$, $p < .01$. Participants with a high school diploma or GED responded significantly faster than participants with some college. There were no significant differences on any condition of any Stroop test based on marital status.

Prior to examining the main hypotheses, two preliminary analyses were undertaken in order to

TABLE 2
Response time means and standard deviations for all conditions of Stroop tests based on education

		GED or high school diploma		*Some college or more education*			
Measure	*Condition*	*Mean*	*SD*	*Mean*	*SD*	*F*	*p*
VRST	WR	911.74	86.41	1,017.22	153.49	5.05	<.05
	CN	1,011.14	117.30	1,114.58	150.87	4.63	<.05
	I	1,183.24	186.70	1,262.05	238.34	1.07	0.31
	CI	1,190.36	151.76	1,314.05	249.25	2.59	0.12
ANAM	WR	612.43	157.68	768.89	145.48	10.58	<.01
	CN	545.11	109.48	667.84	138.56	8.29	<.01
	I	737.97	228.91	853.86	183.43	3.34	0.07
D-KEFS	WR	400.00	62.72	431.67	89.49	1.37	0.25
	CN	533.85	86.17	582.22	117.38	1.84	0.18
	I	940.00	276.53	1,048.89	239.35	1.82	0.18
	CI	1,112.31	174.46	1,223.89	263.85	2.00	0.16

Note. Conditions of Stroop tests: VRST = Virtual Reality Stroop Task; ANAM = Automated Neuropsychological Assessment Metrics; D-KEFS = Delis–Kaplan Executive Function System. WR = Word Reading condition; CN = Color Naming condition; I = Interference condition; CI = Complex Interference condition; *SD* = Standard deviation.

determine any effects of order of test administration on the VRST and whether the VRST was correlated with either of the screening measures (BDI–II and PCL–M). There were no significant differences based on order of administration on the VRST response times. Correlations between the VRST variables and the PCL–M or BDI–II were not significant (see Table 3).

Convergent and discriminant validity

Bivariate correlations were conducted between all Stroop conditions (see Table 3). Based on Cohen's (1988) description of the magnitude of effect sizes, results show that the VRST was moderately correlated with the D-KEFS Stroop test and was highly correlated with the ANAM Stroop test. Although the VRST conditions had significant correlations with the ANAM Procedural Reaction Time and moderate correlations with the ANAM Code Substitution, the ANAM Code Substitution Delayed test showed only low correlations with the VRST word reading and color naming conditions (but no significant correlation with interference and complex interference conditions). The VRST conditions were not correlated with the following ANAM tests: Simple Reaction Time, Math Processing, or Tower Test. The VRST conditions were also not correlated with the any trials of the traditional PASAT, nor were they correlated with any portion (learning and memory trials and PASAT trials) of the Virtual Reality City Memory PASAT Test.

The pattern of scores seen in the VRST was generally similar to that seen on the D-KEFS and ANAM versions of the Stroop (see Figure 1). For each version of the Stroop, response times were collected in milliseconds. There were significant differences in response times according to test conditions (i.e., word reading, color naming, interference, and complex interference; see Table 4). As seen in Table 4, the word reading condition of the VRST and D-KEFS took significantly less time than the corresponding test's color naming condition ($p <$ .05). For the ANAM, this pattern was reversed, and the color naming condition took significantly less time than the word reading condition. For all versions of the Stroop, the interference condition took significantly longer than the word reading and color naming conditions. The complex interference conditions (VRST and D-KEFS) took the longest of all conditions.

DISCUSSION

The current study served as the first to explore the validity of the VRST, an innovative virtual reality Stroop assessment, with a military population. Consistent with predicted hypotheses, the results of the current study establish the preliminary convergent and discriminant validity of the VRST with an active duty military sample.

Tasks performed in a virtual environment may require additional cognitive demands compared to traditional versions of neuropsychological assessments. Additional cognitive demands are thought

TABLE 3
Correlations matrix

	VRST (RT)			
Measure	*WR*	*CN*	*I*	*CI*
VRST (RT)				
Word Reading	1.00	.80**	.80**	.78**
Color Naming	.80**	1.00	.78**	.67**
Interference	.80**	.78**	1.00	.85**
Complex Interference	.78**	.67**	.85**	1.00
BDI-II	.06	−.04	.01	.05
PCL-M	.01	−.08	−.06	−.07
D-KEFS Stroop (RT)				
Word Reading	.25	.04	.25	.18
Color Naming	.42**	.26	.34*	.40**
Interference	.46**	.37*	.49**	.41**
Complex Interference	.25	.24	.32*	.22
ANAM Stroop (RT)				
Word Reading	.75**	.66**	.61**	.63**
Color Naming	.77**	.64**	.61**	.73**
Interference	.63**	.55**	.64**	.67**
ANAM (RT)				
Simple Reaction Time	.26	.19	.11	.15
Simple Reaction Time 2	.14	.16	.20	.20
Procedural Reaction Time	.73**	.64**	.63**	.65**
Code Substitution	.53**	.37*	.32*	.37*
Code Substitution Delayed	.30*	.31*	.16	.21
Matching to Sample	.32*	.39**	.30*	.21
Math Processing	.22	.22	.26	.36*
Tower Test	−.05	.04	−.04	.02
PASAT (# correct)				
Trial 1	−.39	−.24	−.34	−.35*
Trial 2	−.24	−.09	−.09	−.14
Trial 3	−.24	−.12	−.16	−.19
Trial 4	−.32	−.23	−.19	−.15
VR City Memory				
Learning and Memory trials (# correct)				
Trial 1	−.24	−.09	−.02	.03
Trial 2	−.27	−.13	−.10	.00
Trial 3	−.38	−.33	−.17	−.03
Delayed Trial	−.52	−.50	−.35	−.19
PASAT Trials (# correct)				
3-second trial	−.27	−.23	−.23	−.23
2-second trial	−.09	−.10	−.16	−.11

Note. $N = 49$. VRST = Virtual Reality Stroop Task; BDI-II = Beck Depression Inventory–Second Edition; PCL-M = PTSD Checklist–Military version; D-KEFS = Delis-Kaplan Executive Function System; ANAM = Automated Neuropsychological Assessment Metrics; PASAT = Paced Auditory Serial Addition Test; VR = Virtual Reality; RT = response time; WR = Word Reading condition; CN = Color Naming condition; I = Interference condition; CI = Complex Interference condition.
*Correlation is significant at the .05 level (2-tailed); **Correlation is significant at the .01 level (2-tailed).

to be required due to the divided attention necessary in interacting in a virtual environment. Consistent with this hypothesis, mean response times for the VRST in the current study were significantly longer than ANAM and the D-KEFS Stroop tests on all conditions. Although faster response times can reflect a trade-off of speed for accuracy, this did not appear to be the case in this sample. Cognitive tasks performed in a virtual environment may better replicate natural conditions compared to traditional paper-and-pencil tests conducted in typical test settings. Particularly in the military, where questions relating to return of duty are so critical, valid neuropsychological assessments that are specifically designed for use with military populations could provide a particularly useful tool.

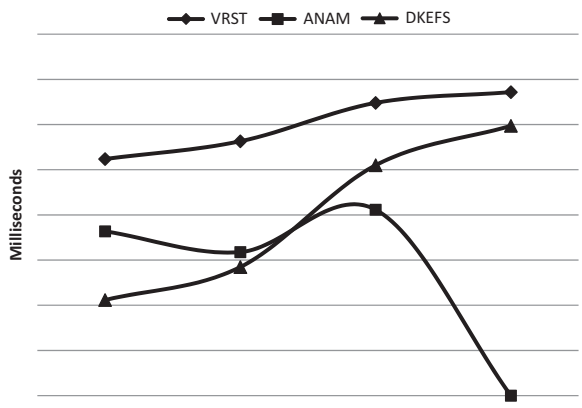

Figure 1. Comparison of the average response times for three versions of the Stroop test (VRST, ANAM, and D-KEFS) (N = 49). VRST = Virtual Reality Stroop Task; ANAM = Automated Neuropsychological Assessment Metrics; D-KEFS = Delis-Kaplan Executive Function System.

There is great need for improved detection of neurocognitive deficits in the military setting, and valid assessments within military-specific VEs may provide benefits over traditional assessments and computerized assessments. Although this study did not assess the criterion validity of the VRST, future research should explore this issue and seek to determine the value of the VRST for military clinical neuropsychology. For example, the Army actively uses structured, observer-scored evaluations of a wide range of common military tasks (e.g., steps for assessing and rendering first aid upon encountering a casualty). Future research could study the relationship of the VRST to skills relevant to military performance. This information could potentially be of value to clinicians faced with difficult decisions regarding return-to-duty for soldiers. VR-based neuropsychological assessments, such as the VRCPAT, would also provide a set of neurocognitive benchmarking tasks, providing the opportunity for use as standard measurements for other military human performance research. Once normative psychometric properties of VR measures of neuropsychological functioning are established, the Army would have a comprehensive, flexible, and scalable VR assessment system that could be extended to investigate a substantial set of pragmatic military performance assessment questions. For example, the neurocognitive components that govern decision making under varying stimulus conditions of cognitive load, fatigue, altitude, and stress induction have the potential to be evaluated with VRCPAT battery and other VR-based neuropsychological assessments.

In the current sample, differences in performance were found based on the demographic variables age and education. Participants who were older tended to respond more slowly for some of the conditions of the VRST, and participants with a high school diploma or GED tended to have faster response times than participants with higher levels of education. Age-based differences were consistent with prior research in the general population, but faster responses for those with less education were unexpected. It may be that a variable unmeasured in the current sample, such as level of physical fitness, may be moderating the relationship found between response time and education. Based on prior research that found faster reaction times among more physically fit individuals (Kramer, Erickson & Colcombe, 2006; Spirduso, 1980), one might speculate that soldiers with less education were more likely to be employed in physically demanding military occupational specialties (e.g., combat arms) requiring higher levels of physical fitness, resulting in faster responses. Race group sample sizes were too small to statistically evaluate differences based on this demographic variable. Future research with larger samples should examine the potential for additional demographic influences on performance.

The conditions of VRST were highly intercorrelated in this sample, which raises the possibility of a single underlying factor such as processing efficiency associated with following a simple set of rules (such as measured by the ANAM subtest Procedural Reaction Time, which all conditions

TABLE 4
Average response time in milliseconds for three versions of the Stroop test

	Response time (milliseconds)					
Measure	*WR*	*CN*	*I*	*CI*	*F*	*p*
VRST	1,047.46 (153.53)	1,125.90 (165.99)	1,295.97 (246.70)	1,343.42 (239.21)	81.96	<.005
ANAM	727.38 (162.85)	635.28 (141.39)	827.61 (203.09)	N/A	56.03	<.005
D-KEFS	423.27 (83.80)	569.39 (111.21)	1,020.00 (251.50)	1,194.29 (246.68)	324.41	<.005

Note. N = 49. VRST = Virtual Reality Stroop Task; ANAM = Automated Neuropsychological Assessment Metrics Stroop Test; D-KEFS = Delis-Kaplan Executive Function System Color–Word Interference Test; WR = Word Reading condition; CN = Color Naming condition; I = Interference condition; CI = Complex Interference condition; N/A = not applicable.

of the VRST correlated highly with). Although the results generally supported the convergent and discriminant validity of the VRST, future research discriminating the individual VRST conditions from additional established tests of psychomotor speed and procedural reaction time would strengthen current findings. In addition, some correlations between the VRST conditions and other tests in the current battery were unexpected. The VRST word reading (WR) condition had a higher correlation with the D-KEFS color naming (CN) and interference (I) than the D-KEFS WR condition. VRST color naming (CN) and complex interference (CI) conditions also had higher correlations with non-parallel subtests of the D-KEFS Color–Word Test. Future research may also benefit from examining these issues using factor analysis and larger sample sizes.

While the results of the current study provide good support for the validity of the VRST, some limitations exist, including the above-described differences in performance based on demographic variables in the current sample, the potential effects of simulator sickness in a virtual environment, the high intercorrelations with different conditions of the VRST, and some unexpected higher magnitude correlations with the VRST conditions and other nonparallel tests in the current battery. Although other tests of cognition have relied on similar sample sizes for psychometric support (Spreen, Strauss, & Sherman, 2006), results of the current study would be strengthened with a larger sample size. Although the VRST may enhance criterion validity relative to traditional tests, this study did not include criterion validity, and there is no current evidence that the VRST has a stronger relationship with relevant outcomes compared to the ANAM, D-KEFS, or other versions of the Stroop. Future studies examining relevant outcomes and scores on the traditional, computerized, and VR-based neuropsychological assessments would help examine the potential value of VRST in this regard.

A limitation of VR-based assessments is the risk of simulator sickness in participants. Simulator sickness occasionally occurs during the use of virtual environments and can include symptoms such as dizziness, headache, sweating, nausea, or vomiting (Kolasinski, 1995). During the current study, three participants terminated testing procedures due to simulator sickness, with symptoms including headaches and nausea. The use of VR assessments in clinical populations with significant prior neurocognitive deficits and medical complications requires careful consideration. In particular, a common symptom of individuals with traumatic brain injury is headaches, and exacerbation of these symptoms could decrease the validity of the test results obtained.

Another limitation in the current study is the lack of adequate variance in accuracy scores to determine whether there was a trade-off between speed and accuracy. For example, D-KEFS percent of items correct ranged from 97% to 99%. This reflects a ceiling effect and may attenuate correlations with scores on other Stroop tests due to restricted range. Future research would benefit from making comparisons between speed and accuracy (if comparisons are possible) to determine whether there is a trade-off between these variables and how demographic variables may be affected.

With the rise of TBI prevalence among service members and the associated risk for persistent neurocognitive deficits, there is a clear need for ongoing research and development of neuropsychological assessment methods. Further, since such injuries can greatly interfere with the service member's ability to perform complex cognitive and emotional processing tasks involved in optimal performance at work, these assessments need to provide clinicians with critical information to inform decision-making regarding return to work questions as well as treatment planning. The current study provides initial promise for neurocognitive assessments using VR by supporting the validation of the VRST in a military sample. Future research should examine the validity of the VRST for predicting real world military performance.

Original manuscript received 20 April 2012
Revised manuscript accepted 10 October 2012
First published online 16 November 2012

REFERENCES

Alexander, M. P. (1995). Mild traumatic brain injury: Pathophysiology, natural history, and clinical management. *Neurology*, *45*, 1253–1260.

Armed Forces Health Surveillance Center (2011). TBI numbers by severity. All armed forces: Department of Defense numbers for traumatic brain injury '00 - '11 Q1 totals. Arlington, VA. http://dvbic.org/TBI-Numbers.aspx

Automated Neuropsychological Assessment Metrics (Version 4) [Computer software]. (2007). Norman, OK: Center for the Study of Human Operator Performance (C-SHOP).

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: The Psychological Corporation.

Belanger H. G., Curtiss, G., Demery, J. A., Lebowitz, B. K., & Vanderploeg, R. D. (2005). Factors moderating neuropsychological outcomes following mild traumatic brain injury: A meta-analysis. *Journal of the International Neuropsychological Society*, *3*, 215–227.

Binder, L. M. (1997). A review of mild head trauma. Part II: Clinical implications. *Journal of Clinical and Experimental Neuropsychology*, *19*, 432–457.

Binet, A., & Simon T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique*, *11*, 191–244.

Bohnen, N., & Jolles, J. (1992). Neurobehavioral aspects of postconcussive symptoms after mild head injury. *Journal of Nervous and Mental Disease*, *180*, 683–692.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (second edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Crawford, J. R., Obonsawin, N. C., & Allan, K. M. (1998). PASAT and components of WAIS–R performance: Convergent and discriminant validity. *Neuropsychological Rehabilitation*, *8*, 255–272.

Deb, S., Lyons, I., Koutzoukis, C., Ali, I., & McCarthy, G. (1999). Rates of psychiatric illness 1 year after traumatic brain injury. *American Journal of Psychiatry*, *156*, 374–378.

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

Department of Defense (2009). *Demographics 2009: Profile of the Military Community*. Arlington, VA: Department of Defense.

Diehr, M. C., Cherner, M., Wolfson, T. J., Miller, W., & Grant, I. (2003). The 50- and 100-item short forms of the Paced Auditory Serial Addition Task (PASAT): Demographically corrected norms and comparisons with the full PASAT in normal and clinical samples. *Journal of Clinical and Experimental Neuropsychology*, *25*, 571–585.

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*, 375–387.

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*, 489–96.

Frencham, K. A., Fox, A. M., & Mayberry, M. T. (2005). Neuropsychological studies of mild traumatic brain injury: A meta-analysis review of research since 1995. *Journal of Clinical and Experimental Neuropsychology*, *27*, 334–351.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills*, *44*, 367–373.

Hofman, P. A., Stapert, S. Z., van Kroonenburgh, M. J. Jolles, J., de Kruijk, J., & Wilmink, J. T. (2001). MR imaging, single-photon emission CT, and neurocognitive performance after mild traumatic brain injury. *American Journal of Neuroradiology*, *22*, 441–449.

Hofman, P. A., Verhey, F. R., Wilmink, J. T., Rozendaal, N., & Jolles, J. (2002). Brain lesions in patients visiting a memory clinic with postconcussional sequelae after mild to moderate brain injury. *Journal of Neuropsychiatry and Clinical Neurosciences*, *14*, 176–184.

Hoge, C. W., McGurk, D., Thomas, J. L., Cox, A. L. Engel, C. C., & Castro, C. A. (2008). Mild traumatic brain injury in U.S. soldiers returning from Iraq. *The New England Journal of Medicine*, *358*, 453–463.

Jones, W., Loe, S., Krach, S., Rager, R., & Jones, H. (2008). Automated Neuropsychological Assessment Metrics (ANAM) and Woodcock–Johnson III tests of cognitive ability: A concurrent validity study. *Clinical Neuropsychologist*, *22*, 305–320.

Kolasinski, E. M. (1995). Simulator Sickness in Virtual Environments. *U.S. Army Research Institute, Simulator Systems Research Unit.* Technical Report 1027.

Kramer, A. F., Erickson, K. I., & Colcombe, S. J. (2006). Exercise, cognition, and the aging brain. *Journal of Applied Physiology*, *101*, 1237–1242.

McLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

McLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General*, *121*, 12–14.

Parsons, T. D. (2011) Neuropsychological assessment using virtual environments: Enhanced assessment technology for improved ecological validity. In S. Brahnam (Ed.), *Advanced computational intelligence paradigms in healthcare: Virtual reality in psychotherapy, rehabilitation, and assessment* (pp. 271–289). Berlin, Germany: Springer-Verlag.

Parsons, T. D., Cosand, L., Courtney, C., Iyer, A., & Rizzo, A. A. (2009). Neurocognitive workload assessment using the Virtual Reality Cognitive Performance Assessment Test. *Lecture Notes in Artificial Intelligence*, *5639*, 243–252.

Parsons, T. D., Courtney, C. G., Arizmendi, B., & Dawson, M. (2011). Virtual Reality Stroop Task for neurocognitive assessment. *Medicine Meets Virtual Reality*, *18*, 433–439.

Parsons, T. D., Courtney, C., Rizzo, A. A., Armstrong, C., Edwards, J., & Reger, G. (2012). Virtual Reality Paced Serial Assessment Test for neuropsychological assessment of a military cohort. *Studies in Health Technology and Informatics*, *173*, 331–337.

Parsons, T. D., & Rizzo, A. A. (2008). Neuropsychological assessment of attentional processing using virtual reality. *Annual Review of CyberTherapy and Telemedicine*, *6*, 23–28.

Parsons, T. D., Rizzo, A. A., Bamattre, J., & Brennan, J. (2007). Virtual reality cognitive performance assessment test. *Annual Review of CyberTherapy and Telemedicine*, *5*, 163–171.

Parsons, T. D., Rizzo, A. A., & Buckwalter, J. G. (2004). Backpropagation and regression: Comparative utility for neuropsychologists. *Journal of Clinical and Experimental Neuropsychology, 26,* 95–104.

Parsons, T. D., Silva, T. M., Pair, J., & Rizzo, A. A. (2008). A virtual environment for assessment of neurocognitive functioning: Virtual reality cognitive performance assessment test. *Studies in Health Technology and Informatics*, *132*, 351–356.

Ponsford, J., Wilmott, C., Rothwell, A., Cameron, P., Kelly, A. M., Nelms, R., Curran, C., & Ng, K. (2000). Factors influencing outcome following mild traumatic brain injury in adults. *Journal of the International Neuropsychological Society*, *6*, 568–579.

Ryan, L. M., & Warden, D. L. (2003). Post concussion syndrome. *International Review of Psychiatry*, *15*, 310–316.

Sbordone, R. J. (1996). Ecological validity: Some critical issues for the neuropsychologist. In: R. J. Sbordone and C. J. Long (Eds.), The ecological validity of

neuropsychological testing, (pp. 15–41). Orlando: GR Press/St. Lucie Press.

Spirduso, W. W. (1980). Physical fitness, aging, and psychomotor speed: A review. *Journal of Gerontology*, *35*, 850–865.

Spreen, O., Strauss, E., & Sherman, E. M. S. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford: Oxford University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Terrio, H., Brenner, L. A., Ivins, B. J., Cho, J. M., Helmick, K., Schwab, K., Scally, K., Bretthauer, R., & Warden, D. (2009). Traumatic brain injury screening: Preliminary findings in a U.S. Army brigade combat team. *Journal of Head Trauma Rehabilitation*, *24*, 14–23.

Vanderploeg, R. D., Curtiss, G., & Belanger, H. G. (2005). Long-term neuropsychological outcomes following mild traumatic brain injury. *Journal of the International Neuropsychological Society*, *11*, 228–236.

Vasterling, J. J., Proctor, S. P., Amoroso, P., Kane, R., Heeren, T., & White, R. F. (2006). Neuropsychological outcomes in Army personnel following deployment to the Iraq war. *Journal of the American Medical Association*, *296*, 519–529.

Vincent, A. S., Bleiberg, J., Yan, S., Ivins, B., Reeves, D. L., Schwab, K., Gilliland, K., Schlegel, R., & Gordon, D. (2008). Reference data from the Automated Traumatic Brain Injury in an active duty military sample. *Military Medicine*, *173*, 836–852.

Weathers, F. W., Huska, J. A., & Keane, T. M. (1991). PTSD Checklist–Military Version (PCL–M) for DSM–IV: National Center for PTSD–Behavioral Science Division.

Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.