OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 35 (2020) 347–357

Archives of CLINICAL NEUROPSYCHOLOGY

Smartphone-Based Measurement of Executive Function in Older Adults with and without HIV

Raeanne C. Moore[1,*], Laura M. Campbell[2], Jeremy D. Delgadillo[3], Emily W. Paolillo[2], Erin E. Sundermann[1], Jason Holden[1], Pierre Schweitzer[4], Robert K. Heaton[1], Joel Swendsen[4,5]

[1]*Department of Psychiatry, University of California, San Diego, 220 Dickinson Street, San Diego, CA 92103, USA*
[2]*San Diego Joint Doctoral Program in Clinical Psychology, San Diego State University/University of California, 6363 Alvarado Court, San Diego, CA 92120, USA*
[3]*Advancing Diversity in Aging Research (ADAR) Program, San Diego State University, 6505 Alvarado Road, Suite 110, San Diego, CA 92120, USA*
[4]*University of Bordeaux, CNRS UMR, Bordeaux 5287, France*
[5]*National Center for Scientific Research, Ecole Pratique des Hautes Etudes, PSL Research University, Paris, France*

*Corresponding author at: HIV Neurobehavioral Research Program, 220 Dickinson Street, Suite B (8231), San Diego, CA 92103, USA. Tel.: 619-543-5378; fax: 619-543-1235. *E-mail address*: r6moore@ucsd.edu

Received 31 July 2019; revised 5 November 2019; Accepted 17 December 2019

Abstract

Objective: To examine feasibility, convergent validity and biases associated with a mobile color-word interference test (mCWIT) among older persons living with HIV (PLHIV).
Method: Over a 14-day period, 58 PLHIV and 32 HIV-uninfected individuals (aged 50–74) completed the mCWIT on smartphones once per day in real-world settings. Participants also completed a comprehensive laboratory-based neuropsychological evaluation.
Results: A high rate of compliance was observed (86%) in the repeated administration of the mCWIT. A practice effect was observed in the overall sample concerning mCWIT subscores, and these learning effects were greater for PLHIV. Stabilization of performance was observed after 6 (HIV+) and 7 days (HIV−) for completion time and after 2 (HIV−) and 3 days (HIV+) for mCWIT errors. A minor fatigue effect was observed in the overall sample which was unassociated with group status. Moderate to strong correlations were found between mCWIT completion time and mCWIT errors with global neurocognition and with all of the individual neurocognitive domains. The strongest associations were with mCWIT completion time and laboratory-based global neurocognition, executive function, and working memory scores.
Conclusions: Cognitive testing administered within the context of a person's daily life provides qualitatively different data than neuropsychological testing completed in clinical settings, and it may constitute a more ecologically valid indicator of cognitive performance than traditional methods. Mobile cognitive testing has potential to help characterize real-time cognitive performance and serve as a complementary assessment tool to traditional methods.

Keywords: Assessment; HIV/AIDS; Executive functions; Practice effects; Reliable change

Introduction

With advancements in mobile technologies, ecological momentary assessment (EMA) has become a practical means of assessing thoughts, feelings, and behaviors in naturalistic environments. EMA provides insights into individual variability and change over time, and it can reduce retrospective recall or state-dependent biases by sampling behavior in real-time (Shiffman, Stone, & Hufford, 2008; Trull & Ebner-Priemer, 2013). Further, EMA allows for the examination of the complex relationships

doi:10.1093/arclin/acz084 Advance Access publication 13 January 2020

between life activities and a variety of psychosocial or environmental risk and protective factors. These same advantages are useful in assessing cognitive performance and may improve ecological validity; however, only a handful of studies to date have utilized mobile cognitive testing and it is not currently integrated into standard clinical care (Moore, Swendsen, & Depp, 2017).

Neuropsychological testing is commonly administered in an environment free from noise, interruptions, and distractions in order to elicit the best possible performance from patients (i.e., their cognitive capacity). However, there is often a disconnect between what a person can do as measured in a controlled environment, and their cognitive performance in the real world. Patients may score within normal ranges in an optimal test setting, but still report cognitive difficulties at work or at home (Lezak, 2012). Self-administered mobile cognitive tests allow for real-time, naturalistic, repeated assessment that may be more ecologically valid compared to traditional assessments completed in an optimal environment. Although mobile cognitive testing is not intended to replace traditional neuropsychological testing, it may serve as a complimentary assessment tool to better understand day-to-day neurocognition (Miller & Barr, 2017). Unlike traditional lab-based testing, mobile assessments can easily be administered multiple times a day so that intra-individual variability in cognitive functioning can be assessed and studied. Additionally, aggregated within-person estimates of cognitive functioning may produce more stable estimates of these variables which promote diagnostic accuracy and provide more sensitive data for the detection of neuroimaging brain markers (Allard et al., 2014).

Impairment in cognition is observed in the growing population of older PLHIV (Heaton et al., 2011; Johns et al., 2012; Kramer et al., 2006; Woods, Moore, Weber, & Grant, 2009). However, only a minority of these individuals has access to cognitive evaluations. Mobile testing of cognition, administered through smartphones and other devices that have now become ubiquitous, holds promise for broadening the reach of neuropsychological testing and thus allowing greater access to "cognitive check-ups." However, given that these recent innovations are still in their infancy, studies examining their feasibility and validity would constitute an important contribution to clinical research, and eventually treatment.

In this study, we evaluated the feasibility, validity, and biases associated with a mobile cognitive task: the mobile color-word interference test (mCWIT). The mCWIT is based on the Stroop paradigm, a neuropsychological test that is commonly used for measuring the capacity of an individual to inhibit a predominant response (Faria, 2015; Strauss, Sherman, & Spreen, 2006). The Stroop demonstrates that individuals typically take longer to respond when instructed to name the color of the ink a word is written in when those words are incongruent with the ink color (e.g., when the word "red" printed in blue ink) relative to trials where individuals are instructed to simply read the words or name colored blocks. Several, but not all, research studies have shown that PLHIV have significantly slower response times on the Stroop color-word interference test compared to persons without HIV infection (Carey et al., 2004; Maki et al., 2018; Woods et al., 2009). Overall, the Stroop is a widely used and validated paradigm to examine cognitive abilities in several clinical populations, including HIV.

To date, two studies examining patients with schizophrenia and patients with substance use disorders have utilized a French-language version of the mCWIT and have demonstrated feasibility and convergent validity in those populations (Bouvard et al., 2018; Dupuy et al., 2018). In this study, we used the mCWIT administered via EMA methods on smartphones over a 2-week period in order to test cognitive status among older PLHIV and HIV− comparison participants. Specifically, the aims of this study were to (a) examine the feasibility of the mCWIT relative to study acceptance by participants and compliance with the multiple daily assessments; (b) assess the magnitude of practice and fatigue effects of the mCWIT, as well as stabilization of mCWIT performance over the study period; (c) examine mCWIT performance by HIV status; and (d) estimate the convergent validity of the mCWIT with three theoretically-related constructs (i.e., a laboratory-assessed measure of color-word interference (Stroop), sociodemographic variables, and composite global and domain-specific cognitive functioning).

Methods

Participants

Fifty-eight HIV-infected (HIV+) and 32 HIV-uninfected (HIV−) individuals participated in this study, conducted at UCSD's HIV Neurobehavioral Research Program (HNRP). Participants were enrolled between February 2016 and October 2018. Sample size was determined via a power analysis to detect a medium effect with 80% power at $p < 0.05$. Inclusion and exclusion criteria were kept to a minimum to increase generalizability. Inclusion criteria were 50 years of age or older, able to provide written informed consent, and fluent in English. Exclusion criteria were history of a non-HIV neurological disorder (e.g., stroke with lasting neurological or neuropsychiatric consequences, head injury with loss of consciousness greater than 30 min or neurological complications, seizure disorder), serious mental illness (e.g., schizophrenia, bipolar disorder), color-blindness, or history of severe learning disability as measured by a WRAT-4 (Wilkinson & Robertson, 2006) score < 70. Participants with a positive alcohol breathalyzer or urine toxicology for illicit substances at the baseline visit (excluding cannabis products) were rescheduled. Laboratory-based data from participants from other HNRP studies were linked if they completed a comprehensive

Table 1. HIV Neurobhehavioral Research Program (HNRP) neuropsychological battery

Speed of information processing	**Working memory**
WAIS-III Digit Symbol	WAIS-III Letter-Number Sequencing
WAIS-III Symbol Search	Paced Auditory Serial Addition Task (PASAT)
Trail Making Test Part A	**Complex motor skills**
Stroop Color and Word Test (color trial)	Grooved Pegboard Test
Learning and memory (two domains)	**Verbal fluency**
Hopkins Verbal Learning Test-Revised (HVLT-R)	Controlled Oral Word Association Test (FAS)
Brief Visuospatial Memory Test-Revised (BVMT-R)	Category Fluency Test ("animals" and "actions")
Executive functions	**Estimated premorbid IQ**
Wisconsin Card Sorting Test (64-item version)	WRAT-4 Reading (alternating Green and Blue versions)
Trail Making Test Part B	**Psychiatric interviews**
Stroop Color and Word Test (Interference score)[a]	Composite International Diagnostic Interview (CIDI)

[a]Raw scores on the Stroop Interference task is the number of words/colors said correctly in 45 s. If a participant makes an error they are corrected in real time.

neuromedical and neurobehavioral visit within the past 6 months. Participants were either recruited from our existing participant pool (i.e., participants who had previously been enrolled in prior studies at our research center, and who agreed to be contacted for future studies), or newly recruited. Newly recruited PLHIV were recruited from a variety of sources that serve adults with HIV in the San Diego area (e.g., community clinics, health care providers), and newly recruited HIV− participants were recruited via flyers posted throughout the community or via presentations by study staff at community organizations. The UCSD Institutional Review Board approved all study procedures prior to protocol implementation. All participants demonstrated capacity to consent (Jeste et al., 2007) and subsequently provided written informed consent.

Measures and Procedure

The study consisted of two in-person visits and a 14-day period of EMA-based mobile cognitive testing between in-person visits. Participants were not co-enrolled in other studies during the study period. Participants were compensated for in-person assessments as well as for each mCWIT test they completed.

Baseline visit: neuromedical and neuropsychological evaluations. At the baseline visit, neuromedical and neuropsychological evaluations were administered to participants, with the exception of participants who previously participated in a different HNRP study within the past 6 months. HIV serostatus was determined using an HIV/HCV antibody point-of-care rapid test and confirmatory western blot analyses for all participants. Participants completed a standardized interview about their medical/drug history. Among HIV+ participants, the following HIV characteristics were obtained: self-reported estimated duration of HIV infection, historical and current use of antiretroviral therapy (ART) and other medications, nadir CD4 T-cell count, and historical AIDS diagnosis. Current CD4 cell count and HIV RNA plasma levels were measured using reverse transcriptase-polymerase chain reaction with a lower limit of quantitation (LLQ) at 50 viral copies/ml (undetectable viral loads characterized as below the LLQ).

Details regarding the widely used standard HNRP neuropsychological test battery have been previously published (Heaton et al., 2010). This battery assesses cognition in the following domains: estimated premorbid verbal intelligence (IQ), verbal fluency, executive functioning, speed of information processing, verbal and visual learning, delayed recall memory, attention/working memory, and complex motor skills (Cysique et al., 2011). A summary of the battery is provided in Table 1. The HNRP Core battery and standardized interpretation algorithms were constructed in accordance with the Frascati guidelines (Antinori et al., 2007). Raw scores from the neuropsychological tests are converted to scaled scores and demographically adjusted T-scores ($M = 50$, $SD = 10$ in healthy subjects; Heaton et al., 2004; Heaton, Taylor, & Manly, 2002). Practice effect algorithms are applied to control for prior exposure to neuropsychological testing. See Cysique et al. (2011) for details of our practice effect algorithms. We chose to use raw scores from the standard laboratory-based Stroop Color Word Interference Test (Golden, Hammeke, & Purisch, 1978) for direct comparison and scaled scores from each neurocognitive domain in our validity analyses, as our mCWIT data is not demographically adjusted.

During their baseline visit, participants were provided a Samsung Galaxy S 4.2 YP-GI1 8GB smartphone and a smartphone operating manual. The Galaxy Player 4.2 has a 4.2″ IPS display at 800 × 480, 1 GHz processor, front and rear cameras, and Android 2.3 Gingerbread. Android version 2.3.6 was installed. The smartphone was provided with the purpose of providing participants with a standard platform that would minimize a bias towards participants who do not own smartphones. Participants completed a 20–30-min training session to become familiar with the structure and format of the smartphone and how to complete

the daily surveys and the mCWIT. To ensure security of the data, the study phone's operating system was encrypted in case the smartphone was lost or stolen.

Smartphone monitoring: 14-day EMA and mobile cognitive testing. Participants received four surveys per day on the study smartphones for 14 days. Each survey took approximately 3 min to complete. The timing of surveys was adjusted to accommodate each participant's preferred sleep–wake schedules, and the smartphones delivered surveys at random intervals approximately 3 hr apart throughout the day. Each of the four EMA surveys asked questions about the participant's daily functioning, including where they were, who they were with, and what they were doing, current mood, substance use, and socialization (Moore et al., 2016; Paolillo, Obermeit, et al., 2018a; Paolillo, Tang, et al., 2018b).

A sample screenshot of the mCWIT is presented in Fig. 1. Our version of the mCWIT is the same as the French version, just translated to English. The order of color words and pairings of incongruent color-words differed at each administration. The timing of administration was counter-balanced to ensure that the mobile cognitive test was administered at different times of the day. Modeled after the interference trial of the Stroop Color and Word Test, the mCWIT provides participants with a list of 16 color words (four instances each of YELLOW, RED, BLUE, GREEN, in random order) in different colors (also yellow, red, blue, green). All words were written in a color that was incongruent to the meaning of the written word (e.g., the word "Blue" would appear in the color of red, yellow, and green, but will not appear in the color blue). The order of words and colors was randomized, and each word and color appeared once per line. Once prompted to start the mCWIT, participants were instructed to not read the words but to say the colors in which they are written aloud and as quickly as possible. Participants were provided a maximum of 60 s to complete the task, with a timer provided at the bottom of the screen. Once they completed the task, participants had the option to select "Done" on the screen or let the time limit expire to finish the test. Responses were audio-recorded on the study smartphones, and each audio file was listened to and scored independently by two trained raters. The number of correct responses, number of errors, time to completion, and potential cheating (e.g., someone else who was not the participant completing the test) were rated. Potential cheating was determined if a different voice from the participant's voice was heard on the recording (e.g. from one participant, "you just say the color of the ink, don't read the word," then another voice was heard completing the task). If a participant finished before 60 s and starting saying the colors over again from the beginning, these additional responses were not recorded. Trials in which <15 of the 16 words on the list were completed were considered invalid due to incomplete data, and thus excluded. Only 17 trials had <15 responses and were discarded, equating to $\sim$1.5% of the total number of trials. If they completed 15 words and missed the 16th word, the last word missing was coded as an error. Time to completion was scored from the time the recording began until the participant completed stating the colors. Self-corrected responses were scored as correct. If there were discrepancies in ratings between the two raters (JDD; LMC), a third independent rater would listen to and score the audio file. There was minimal judgment required, especially for the time scores, because the recordings were time-stamped. The percent agreement for mCWIT time scores (+/− 2 s) was 98.42%. The percent agreement for mCWIT error scores was 91.67%. All discrepancy scores were also reviewed by the senior author (RCM).

Follow-up visit: feedback survey. After the 14-day mobile cognitive testing period, participants returned the smartphones and completed a post-study feedback questionnaire to assess acceptability of study procedures. They also were provided with an opportunity to describe their overall experience and challenges they encountered.

Statistical Analyses

Data were analyzed using SPSS version 26 (IBM Corp., 2013) and R software (R Core Team, 2015). There were two HIV− participants who did not have any correct responses on all instances of the mCWIT; these two participants were removed from analyses, as their data indicates they likely did not understand or faithfully comply with the task instructions. There were also one HIV+ participant and one HIV− participant who had 14% mCWIT compliance (the minimal compliance threshold of 30% was set for analyses) and were thus removed from analyses. HIV serostatus group differences on demographic factors, neurocognitive performance, and mCWIT (percent compliance, mean errors, and mean response time over a 14-day period) were assessed via chi-Square, Fisher's exact tests, and *t*-tests (or non-parametric equivalent), as appropriate. Analyses of EMA fatigue and practice effects as well as group differences in these effects were examined by random coefficient linear models for continuous outcomes and Bernoulli models for dichotomous outcomes. Fatigue effects were defined as decreases in the number of mCWIT completed over the 2-week period, and practice effects were defined as improved changes in test performance over the 2-week period. Two spline regression models (one for HIV+ and one for HIV− participants) were conducted to determine whether there was a point (day on study) at which mCWIT performance maximizes and stabilizes. In order to examine convergent validity,

Fig. 1. Sample screenshot of the mobile color-word interference test (mCWIT).

the number of mCWIT errors and time to completion were correlated with in-person raw Stroop scores and neuropsychological domain uncorrected scaled scores. Statistical differences between these dependent correlations were examined using established methodology (Steiger, 1980).

Results

Acceptability, Feasibility, and Compliance

Group demographics, clinical characteristics, and cognitive performance scores are presented in Table 2. Although we provided smartphones to participants for this study, we also assessed their personal smartphone ownership and usage. Overall, 95% of the participants owned personal smartphones, 97% of these reported carrying their smartphone on a regular basis, and 75% of the participants used their smartphone between 0 and 4 hr per day, with the remaining 25% reporting more use. Participants reported no difficulty carrying both their personal and study smartphone for the 2-week study period. Participants had a high rate of compliance with the mCWIT, with both the HIV+ and HIV− participants completing an average of 86% of the possible assessments in their daily lives, resulting in 1,100 observations of the mCWIT. In the overall sample, number of assessments completed did not significantly differ between persons who were employed ($n = 29$; $M = 84.0\%$; $SD = 16.6\%$) compared to persons who were retired or unemployed ($n = 60$; $M = 87.1\%$; $SD = 14.0\%$; $t = 0.93$, $p = 0.35$). Compliance examination by our raters identified six observations of suspected cheating, defined as the participant possibly receiving help from others or having others complete the mCWIT on their behalf. These six observations were removed from analyses.

Practice and Fatigue Effects

In the overall sample, a practice effect from repeated testing was observed. The errors (coefficient: −0.11, $SE = 0.02$, $T = -6.68$, $df = 1046$, $p < 0.001$) and completion time (coefficient: −0.72, $SE = 0.04$, $T = -16.27$, $df = 1044$, $p < 0.001$) on the mCWIT decreased with practice (as a function of day of study). An effect of HIV status was observed in that PLHIV demonstrated practice effects of greater magnitude than HIV− participants relative to mCWIT errors, as PLHIV had more errors than HIV− participants on early mCWIT trials. There was also a very minor but significant fatigue effect, with slight increases in the likelihood of a non-completed mCWIT tests as a function of day in the study (OR = 1.085, 95% CIs = 1.039 to 1.132, $p < 0.001$). The minor fatigue effects did not vary by HIV status (OR = 0.992, 95% CIs = 0.908 to 1.080, $p = 0.86$).

Table 2. Demographics and clinical characteristics

	HIV+($n = 58$)	HIV−($n = 32$)	Test-statistic[a]	p value
Age (years)	59 (6.3) Range = 50–73	59 (6.7) Range = 50–74	0.01	0.99
Sex (male)	49 (84%)	17 (53%)	10.37	<0.01
Race/ethnicity (White)	39 (66%)	22 (67%)	0.87	0.83
Education (years)	14.2 (2.6)	14.9 (2.5)	1.19	0.24
Employment status (employed)[b]	17 (29.8)	12 (37.5)	0.55	0.46
WRAT-Reading	103.1 (14.6)	106.3 (16.0)	0.96	0.34
Smartphone ownership (iPhone or Android vs. no phone or other)	51 (88%)	26 (81%)	0.75	0.39
History of AIDS	39 (67%)	—	—	—
Current CD4 count[c]	706.5 [549.5, 871.5]	—	—	—
Nadir CD4 count[c]	173.0 [43.5, 300]	—	—	—
Estimated duration of infection (years)	23.3 (7.3)	—	—	—
On antiretroviral therapy	55 (95%)	—	—	—
GDS-Impaired	22 (38%)	11 (34%)	0.11	0.74
Global Cognition Scaled Score (SS)	8.7 (2.0)	9.4 (2.0)	1.53	0.13
Verbal fluency SS	10.3 (2.7)	11.5 (2.8)	1.88	0.06
Executive functioning SS	8.4 (2.5)	9.0 (2.3)	1.18	0.24
Processing speed SS	9.5 (2.4)	10.1 (2.7)	1.20	0.24
Learning SS	6.9 (2.7)	7.8 (2.6)	1.60	0.11
Delayed recall SS	7.0 (2.4)	7.9 (2.8)	1.46	0.15
Working memory SS	10.0 (2.8)	10.2 (2.9)	0.34	0.73
Complex motor skills SS	7.7 (2.8)	7.9 (2.4)	0.26	0.80
In-person color-word interference test raw-score (number correct in 45 s)	36.4 (12.2)	38.6 (9.4)	0.89	0.38
Percent compliance (average number of mCWIT completed out of 14 possible trials)[c]	85.0 [14.29–100]	84.2 [14.29–100]	0.05	0.95
mCWIT average response time	23.70 (6.1) [14.8–46.9]	21.7 (4.3) [14.7–31.6]	1.68	0.10
mCWIT average errors	0.94 (1.08) [0.0–4.5]	0.40 (0.79) [0.0–3.6]	2.76	<0.01

Note: Values are presented as mean (SD), [b]median [IQR], or *N* (%) WRAT-Reading = Wide Range Achievement Test, Reading subtest GDS = Global Deficit Score

[a]T-tests for continuous variables; Chi[2] for dichotomous variables
[b]$n = 89$
[c]Median [IQR]

Stabilization of mCWIT Performance

As seen in Figure 2, Spline regression models indicated that in the HIV+ group, mCWIT errors decreased significantly over time until study Day 3 (regression of mCWIT errors on days 1–3; $b = -1.34$, $p < 0.001$), after which performance stabilized (regression of mCWIT errors on Days 3–14; $b = -0.04$, $p = 0.16$). In the HIV− group, mCWIT error decreased significantly over time until study Day 2 ($b = -0.80$, $p = 0.04$), after which performance stabilized ($b = -0.03$, $p = 0.32$). Additional spline regression models indicated that in the HIV+ group, mCWIT time performance improved (i.e., faster time to mCWIT completion) significantly over time until study day 6 ($b = -1.57$, $p < 0.001$), after which performance stabilized ($b = -0.14$, $p = 0.27$). In the HIV− group, mCWIT performance improved significantly over time until study Day 7 ($b = -1.52$, $p < 0.001$), after which performance stabilized ($b = -0.17$, $p = 0.28$).

mCWIT Performance by HIV Status

First we examined HIV group differences on average mCWIT performance across the 14 days. Average mCWIT time scores were normally distributed in each group. Average mCWIT error scores were right skewed in each group, as is expected of error scores. Results revealed the HIV+ group completed more errors ($t = 2.76$, $p < 0.01$); this result remained when controlling for demographic characteristics (age, sex, race/ethnicity, and education). The two groups did not differ on average time to complete the task ($t = 1.68$, $p = 0.10$). Because of practice effects and stabilization in performance over time, we compared HIV groups' average mCWIT error performances separately by early trials (study Days 1–3) and late trials (study days 4–14) based on results of spline regressions in the entire sample. HIV group differences were only statistically significant for the early trials (HIV+: $M = 2.41$, $SD = 3.22$; HIV−: $M = 0.80$, $SD = 1.69$; $p = 0.010$). Across later trials, average error performance did not statistically differ by HIV status (HIV+: $M = 0.52$, $SD = 0.83$; HIV−: $M = 0.28$, $SD = 0.61$; $p = 0.149$).

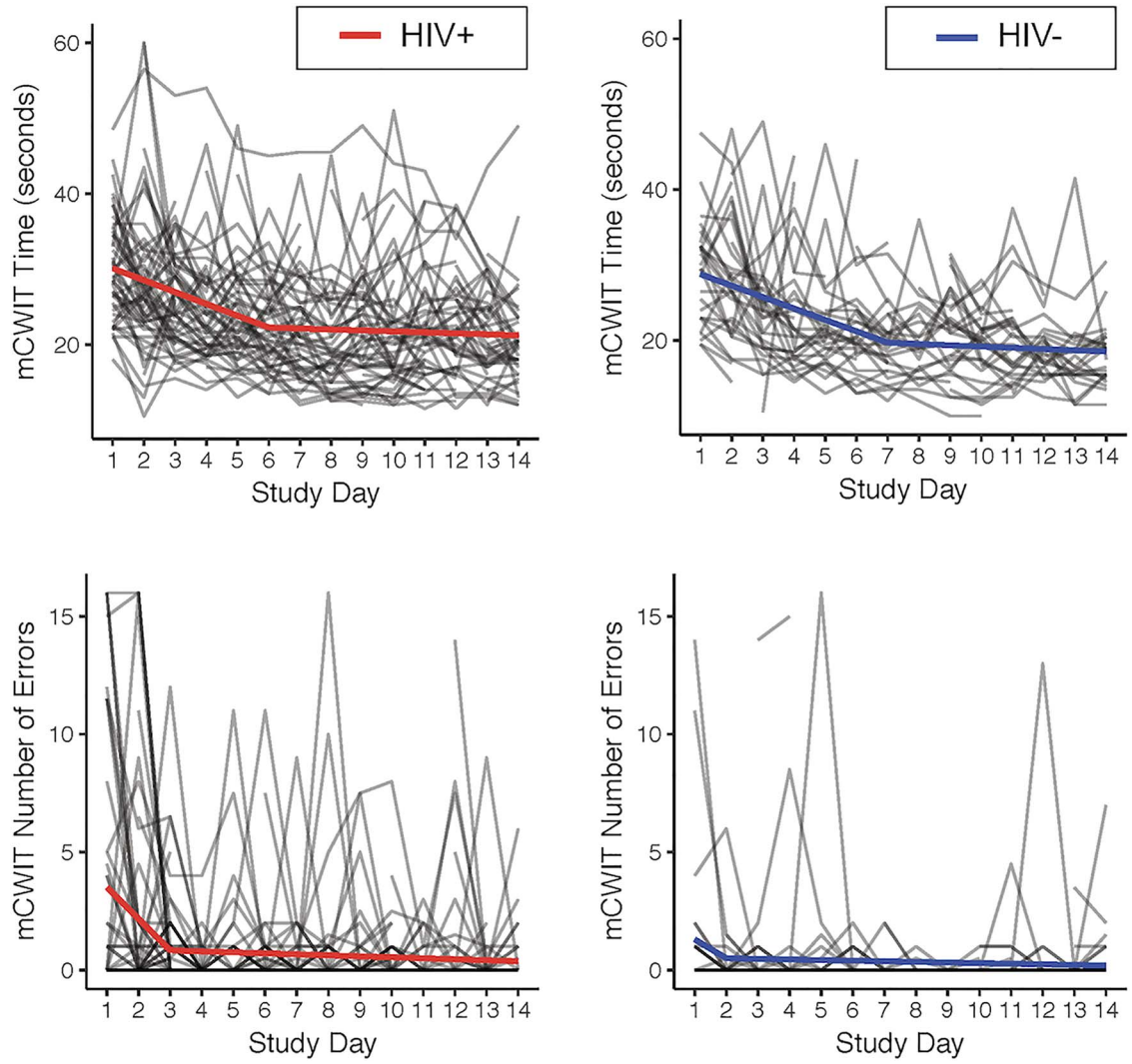

Fig. 2. Stabilization of mobile color-word interference test (mCWIT) performance over time.

Convergent Validity

The mCWIT variables used to examine convergent validity included number of errors and time to complete the task. These two variables were correlated in the overall sample, with longer times correlated with more errors ($r = 0.41$, $p = 0.001$). Fewer errors and shorter mCWIT time was correlated with better in-person (in-laboratory) raw Stroop interference trial times (errors: $r = -0.26$, $p < 0.02$; time: $r = -0.63$, $p < 0.001$; please note higher in-lab Stroop scores indicate better performance, whereas higher mCWIT scores indicate worse performance). mCWIT variables were unrelated to age (errors: $r = 0.00$, $p = 0.97$; time: $r = 0.15$, $p = 0.17$) and sex (errors: $r = 0.15$, $p = 0.16$; time: $r = 0.04$, $p = 0.69$); however, more years of education was related to fewer errors ($r = -0.29$, $p < 0.01$) and faster mCWIT time ($r = -0.22$, $p = 0.04$). In addition, race/ethnicity was related to mCWIT errors ($r = 0.40$, $p < 0.001$) and time ($r = 0.41$, $p < 0.001$). Among PLHIV, HIV disease characteristics (estimated duration of infection, nadir CD4, current CD4, and historical AIDS diagnosis) were unrelated to mCWIT scores.

Next, we examined the correlations between laboratory-based neuropsychological performance, by domain, with the mCWIT. These results are presented in Table 3. More mCWIT errors was related to worse global neurocognition and worse performance in all the individual cognitive domains. For the mCWIT time variable, there were significant correlations with global neurocognition and all of the individual neurocognitive domains. mCWIT time showed the strongest correlation with global neurocognition, which was statistically greater than the correlations between mCWIT and language ($p < 0.001$), learning ($p < 0.001$), and

Table 3. Correlations between the mCWIT, in-lab Stroop, and neuropsychological global and domain performance, including correlations of early and late mCWIT trials with cognitive domains

Laboratory-based neuropsychological performance[a]	**A** In-lab Stroop Color-Word Inhibition Trial (raw)[c]	**B** mCWIT Time[b]	**C** mCWIT # errors	**D** mCWIT time Days 1–7	**E** mCWIT time Days 8–14	**F** mCWIT errors Days 1–3	**G** mCWIT errors Days 4–14
Global performance	0.76[d]**	−0.59**	−0.42**	−0.58**	−0.53**	−0.23*	−0.47**
Language	0.56**	−0.34**	−0.34**	−0.33**	−0.31**	−0.22*	−0.34**
Executive functions	0.76[e]**	−0.55**	−0.22*	−0.55**	−0.47**	−0.11	−0.31**
Processing speed	0.59**	−0.49**	−0.38**	−0.49**	−0.43**	−0.20	−0.44**
Learning	0.54**	−0.36**	−0.36**	−0.36**	−0.31**	−0.20	−0.36**
Recall	0.49**	−0.47**	−0.36**	−0.48**	−0.40**	−0.25*	−0.35**
Working memory	0.60**	−0.53**	−0.30**	−0.50**	−0.48**	−0.09	−0.41**
Complex motor skills	0.45**	−0.41**	−0.23*	−0.38**	−0.41**	−0.14	−0.27*
In-lab Stroop Color-Word Inhibition Trial (raw)	—	−0.63**	−0.26*	−0.62**	−0.58**	−0.10	−0.35**

mCWIT = mobile color-word interference test Columns B & C: Average mCWIT performance over all trials; Columns D & E: early and late trials of mCWIT time; Columns F & G: early and late trials of mCWIT errors $*p < 0.05$; $**p < 0.01$

[a]Uncorrected scaled scores; [b]higher scores = slower (worse) performance; [c]higher scores = faster (better) performance; [d]test included in Global Performance score; [e]test included in Executive Functions domain score

complex motor skills ($p = 0.026$). mCWIT time was also strongly correlated with executive functions, which was statistically greater than the correlations between mCWIT and language ($p = 0.018$) and learning ($p = 0.034$).

We then examined changes in associations between cognitive domains with performance on early trials versus late trials, as our stabilization analyses indicated test performance plateaued early on. We defined early versus late trials by the points at which performance plateaued for errors (Day 3 for entire sample) and for time (Day 7 for entire sample), which were determined in our spline regression models. Thus, we examined the correlations between the cognitive domains and each of the following: (a) early trial mCWIT errors (average errors on Days 1–3); (b) late trial mCWIT errors (average errors on Days 4–14); (c) early trial mCWIT time (average time on Days 1–7); and (d) late trial mCWIT time (average time on Days 8–14). Results are presented in Table 3. There does not appear to be a large difference in the correlations of early versus late mCWIT time performance with cognitive domains (columns F & G). The late trials of mCWIT error performance, however, appear to be much more strongly correlated to each cognitive domain than the early trials of mCWIT error performance (columns D & E).

Discussion

We found support for the feasibility and acceptability of the mCWIT among persons with and without HIV. Participants completed 86% of the mobile cognitive tests, and compliance rates did not differ by HIV status. A minor fatigue effect was observed, with slightly more tests being completed near the beginning of the 14-day study period than near the end. The minor fatigue effects did not vary by HIV serostatus. Practice effects were also observed in the overall sample, and these effects did differ by HIV status. Specifically, both PLHIV and the HIV- controls improved in accuracy with repeated testing and required less time to complete the test over the first 6 to 7 days of the study. Repeated mobile testing of cognitive functions commonly show practice effects, but statistical adjustment for the number of times a test has been previously administered typically does not change the direction or significance of effects (Schweitzer et al., 2016). For a systematic review of the feasibility and psychometric properties of other mobile cognitive tests administered within an EMA context, and for comparison purposes with our results, please see Moore et al. (2017).

A strong correlation was found between our traditional raw Stroop interference trial time and mCWIT performance, and moderate-to-strong correlations were found between our traditional, laboratory-based neuropsychological test scores and those of the mCWIT. A pattern emerged where we found descending correlations with mCWIT time starting with the in-lab Stroop ($r = 0.63$), Global Cognition ($r = 0.59$), Executive Function ($r = 0.55$), Working Memory ($r = 0.53$), Processing Speed ($r = 0.49$), Recall ($r = 0.47$), Motor ($r = 0.41$), Learning ($r = 0.36$) and Language ($r = 0.34$). Although the mCWIT time was most strongly related to global cognition, this would be expected given the "g" effect which will cause all cognitive tests to have some correlation with each other, there does appear to be some specificity to the correlations with mCWIT time. As can be seen, the largest domain-level corrections are with executive function and working memory (which is theoretically strongly related to executive function (McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010). The raw in-person Stroop interference trial

was also related to global neurocognition and every cognitive domain, providing evidence of convergent validity. We also found that later trials of the mCWIT, after error performance improves and stabilizes, are more representative of a person's laboratory-based cognitive performance compared to early trials. This finding further supports the repeated-measures methodology of our mobile assessment.

It is important to highlight that we did observe some education and race/ethnicity effects in regard to mCWIT performance. Older PLHIV who had more years of education made fewer errors, and racial/ethnic group differences in performance were observed. Conversely, we did not observe significant age effects on mCWIT performance. Whereas there is a large body of literature demonstrating age effects on Stroop performance, the majority of these studies find this effect across the lifespan or when comparing younger and older adults (Spieler, Balota, & Faust, 1996; Uttl & Graf, 1997). Therefore, we likely did not observe a significant age effect due to the restricted age range (i.e., age 50–74). It may be possible to develop normative data for mobile cognitive tests if large enough sample sizes can be recruited. We believe a strength of this novel tool is the ability for people to take these tests in real-world setting as they are going about their daily lives and involved in multiple tasks or experiencing diverse distractions. Another strength of mobile cognitive testing is the ability to examine day-to-day variability in cognitive performance and how it relates to real-world outcomes or cognitive decline. Therefore, whereas longitudinal normative data may not be necessary for mobile cognitive tests because each individual is his/her own control, demographic characteristics such as race/ethnicity should be considered when interpreting cross-sectional results.

This study examines a mobile cognitive assessment method that has the potential to provide highly innovative tools for the measurement of cognitive performance in the wild; however, limitations must be addressed to enhance future research in this field. One concern for all mobile cognitive testing is that "cheating" or "help from others" cannot be fully controlled. We were able to identify six instances of suspected cheating with the help of our audio-recorded response format; video recordings would be better able to accurately identify when someone other than the participant completes the assessments, but may also raise more concerns about invasion of privacy. In any case, the number of suspected cheating events represent less than 1% of all collected observations and therefore may be considered as a negligible bias. Additionally, we do not know whether participants were interrupted during the mCWIT task or the degree to which they were invested or motivated in following the testing procedures. These are factors that have potential to invalidate performance on traditional in-person neuropsychological assessments, and as the mCWIT lacks behavioral observations or embedded performance validity tasks; efforts could be made to develop and test built-in measures of invalidity in mobile cognitive tasks (e.g., cutoffs for error rates, random variability in time scores, increasing time scores within sessions). Conversely, an advantage of mobile cognitive tests is the ability to average scores from brief and repeated testing rather than a "one shot" assessment, and therefore these mobile tests may provide a more reliable measurement of cognitive performance and ability despite the aforementioned limitations. The scalability of the mCWIT in its current form is also a concern as the need to have multiple raters limits the sample size and test frequency that could be achieved. However, when we were designing the mCWIT we decided to have participants respond verbally instead of tapping their responses for the following reasons: (a) the ability to assess whether a participant is cheating (e.g., whether someone else taking the test for them), (b) same response paradigm as in-laboratory testing, (c) elimination of a motor component (to help remove bias from motor speed); and (d) responds are not compromised by touch-response delay differences across operating systems and devices. Additionally, each trial takes approximately 1 min to score, which is comparable to scoring other in-lab neuropsychological tests and much less time than administering and scoring a full in-lab neuropsychological battery. Hopefully in the near future, voice recognition software will have improved to the point of detecting accents, various dialects, differentiating between voices, and identifying variations in speech patterns more accurately. Computer-based studies of voice-activated reaction time versions of the Stroop have found shorter reaction times to verbal responses when directly compared to manual (key pressed) responses, demonstrating the utility of voice-activated reaction time scoring in providing a finer grained index of performance (e.g., Pilli, Naidu, Pingali, Shobha, & Reddy, 2013; Repovš, 2004). In addition, we have recently developed an alternative version of the mCWIT which uses touch-response only, includes congruent and interference trials, and has automatic scoring, and we will be able to compare the psychometric properties of the two versions of the task in future work. We will also be able to capture reaction time metrics in our alternate mCWIT. Evidence is emerging that the addition of reaction time testing to traditional neuropsychological assessment increases the sensitivity of cognitive dysfunction detection. A study of cART-treated PLHIV demonstrated that reaction time latency and variability were worse in PLHIV compared to the normative mean, and were also associated with worse global cognitive performance (Ettenhofer et al., 2010). Another point to consider is that we did not utilize mobile versions of word reading and color naming trials that traditionally precede the Stroop interference task. Although this decision was made to maximize compliance and feasibility in a real-world setting, it does perhaps reduce the prepotent response. However, given the following: (1) both the in-lab Stroop and mCWIT scores were comparable, as we used raw data from the in-lab Stroop time score that were only for the interference trial and not corrected for demographics, (2) the in-lab Stroop and mCWIT time were highly correlated, and (3) the in-lab "interference" formula yields a score that has the disadvantage of having less test-retest reliability than the time score for the

interference trial, we believe providing only the interference trial of the mCWIT is an appropriate approach for mobile cognitive testing. Finally, future validation of mobile cognitive testing among this population is needed to further understand clinical relevance.

Overall, we found evidence for the feasibility and initial convergent validity of the mCWIT among PLHIV. The implementation of mobile cognitive testing into research and clinical practice opens the door for frequent, longitudinal tracking of cognitive performance which could improve the early identification of cognitive decline. It could also empower individuals to monitor their own cognition over time, allow patients living in rural environments to have access to cognitive assessments and to enhance the cognitive health dialogue between doctors and patients. This is the third study to validate the mCWIT, as previous work in other populations found it to be a valid assessment tool among healthy controls, patients with substance use disorders and people with schizophrenia (A Bouvard et al., 2018; Dupuy et al., 2018). Future work is needed to continue the validation of the mCWIT as well as additional novel mobile cognitive tests in various clinical populations.

Funding

This work was supported by the National Institutes of Health [grant numbers NIMH K23MH105297, NIMH K23MH107260 S1, and NIMH R21MH116104 to R.C.M., NIDA T32DA031098 to L.M.C., NIA R25AG043364 to J.D.D., and NIAAA F31AA027198 to E.W.P.].

The HIV Neurobehavioral Research Center (HNRC) is supported by Center award P30MH062512 from NIMH.

Conflicts of Interest

R.C.M. is a co-founder of KeyWise AI.

Acknowledgments

*The San Diego HIV Neurobehavioral Research Center [HNRC] group is affiliated with the University of California, San Diego, the Naval Hospital, San Diego, and the Veterans Affairs San Diego Healthcare System, and includes the following: Director: Robert K. Heaton, Ph.D., Co-Director: Igor Grant, M.D.; Associate Directors: J. Hampton Atkinson, M.D., Ronald J. Ellis, M.D., Ph.D., and Scott Letendre, M.D.; Center Manager: Thomas D. Marcotte, Ph.D.; Jennifer Marquie-Beck, M.P.H.; Melanie Sherman; Neuromedical *Component*: Ronald J. Ellis, M.D., Ph.D. (P.I.), Scott Letendre, M.D., J. Allen McCutchan, M.D., Brookie Best, Pharm.D., Rachel Schrier, Ph.D., Debra Rosario, M.P.H.; *Neurobehavioral Component*: Robert K. Heaton, Ph.D. (P.I.), J. Hampton Atkinson, M.D., Steven Paul Woods, Psy.D., Thomas D. Marcotte, Ph.D., Mariana Cherner, Ph.D., David J. Moore, Ph.D., Matthew Dawson; Neuroimaging *Component*: Christine Fennema-Notestine, Ph.D. (P.I.), Monte S. Buchsbaum, M.D., John Hesselink, M.D., Sarah L. Archibald, M.A., Gregory Brown, Ph.D., Richard Buxton, Ph.D., Anders Dale, Ph.D., Thomas Liu, Ph.D.; Neurobiology *Component*: Eliezer Masliah, M.D. (P.I.), Cristian Achim, M.D., Ph.D.; *Neurovirology Component*: David M. Smith, M.D. (P.I.), Douglas Richman, M.D.; *International Component*: J. Allen McCutchan, M.D., (P.I.), Mariana Cherner, Ph.D.; *Developmental Component*: Cristian Achim, M.D., Ph.D.; (P.I.), Stuart Lipton, M.D., Ph.D.; *Participant Accrual and Retention Unit*: J. Hampton Atkinson, M.D. (P.I.), Jennifer Marquie-Beck, M.P.H.; *Data Management and Information Systems Unit*: Anthony C. Gamst, Ph.D. (P.I.), Clint Cushman; *Statistics Unit*: Ian Abramson, Ph.D. (P.I.), Florin Vaida, Ph.D. (Co-PI), Reena Deutsch, Ph.D., Anya Umlauf, M.S.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

References

Allard, M., Husky, M., Catheline, G., Pelletier, A., Dilharreguy, B., Amieva, H. et al. (2014). Mobile technologies in the early detection of cognitive decline. *PLoS One*, *9*(*12*), e112197. doi: 10.1371/journal.pone.0112197.

Antinori, A., Arendt, G., Becker, J. T., Brew, B. J., Byrd, D. A., Cherner, M. et al. (2007). Updated research nosology for HIV-associated neurocognitive disorders. *Neurology*, *69*(*18*), 1789–1799. doi: 10.1212/01.WNL.0000287431.88658.8b.

Bouvard, A., Dupuy, M., Schweitzer, P., Revranche, M., Fatseas, M., Serre, F. et al. (2018). Feasibility and validity of mobile cognitive testing in patients with substance use disorders and healthy controls. *American Journal on Addictions*, *27*(*7*), 553–556.

Carey, C. L., Woods, S. P., Gonzalez, R., Conover, E., Marcotte, T. D., Grant, I. et al. (2004). Predictive validity of global deficit scores in detecting neuropsychological impairment in HIV infection. *Journal of Clinical and Experimental Neuropsychology*, *26*(*3*), 307–319.

Cysique, L. A., Franklin, D., Abramson, I., Ellis, R. J., Letendre, S., Collier, A. et al. (2011). Normative data and validation of a regression based summary score for assessing meaningful neuropsychological change. *Journal of Clinical and Experimental Neuropsychology*, *33*(*5*), 505–522.

Dupuy, M., Misdrahi, D., N'Kaoua, B., Tessier, A., Bouvard, A., Schweitzer, P. et al. (2018). Mobile cognitive testing in patients with schizophrenia: A controlled study of feasibility and validity. *Journal de Thérapie Comportementale et Cognitive*.

Ettenhofer, M. L., Foley, J., Behdin, N., Levine, A. J., Castellon, S. A., & Hinkin, C. H. (2010). Reaction time variability in HIV-positive individuals. *Archives of Clinincal Neuropsychology*, *25*(*8*), 791–798. doi: 10.1093/arclin/acq064.

Faria, C. A., Alves, H. V. D., & Charchat-Fichman, H. (2015). The most frequently used tests for assessing executive functions in aging. *Dementia and Neuropsychologia*, *9*, 149–155. doi: 10.1590/1980-57642015DN92000009.

Golden, C. J., Hammeke, T. A., & Purisch, A. D. (1978). Diagnostic validity of a standardized neuropsychological battery derived from Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology*, *46*(*6*), 1258–1265.

Heaton, R. K., Clifford, D. B., Franklin, D. R., Jr., Woods, S. P., Ake, C., Vaida, F. et al. (2010). HIV-associated neurocognitive disorders persist in the era of potent antiretroviral therapy: CHARTER Study. *Neurology*, *75*(*23*), 2087–2096. doi: 10.1212/WNL.0b013e318200d727.

Heaton, R. K., Franklin, D. R., Ellis, R. J., McCutchan, J. A., Letendre, S. L., LeBlanc, S. et al. (2011). HIV-associated neurocognitive disorders before and during the era of combination antiretroviral therapy: Differences in rates, nature, and predictors. *Journal of Neurovirology*, *17*, 3–16.

Heaton, R. K., Marcotte, T. D., Mindt, M. R., Sadek, J., Moore, D. J., Bentley, H. et al. (2004). The impact of HIV-associated neuropsychological impairment on everyday functioning. *Journal of the International Neuropsychological Society*, *10*(*3*), 317–331. doi: 10.1017/s1355617704102130.

Heaton, R. K., Taylor, M., & Manly, J. (2002). Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. In Tulsky, D. S., Saklofske, D., Heaton, R. K., Chelune, G., Ivnik, R., Bornstein, R. A. et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III*. San Diego: Academic.

I.B.M. Corp. (2013). *IBM SPSS Statistics for Windows (Version 26)*. Armonk, NY: IBM Corp.

Jeste, D. V., Palmer, B. W., Appelbaum, P. S., Golshan, S., Glorioso, D., Dunn, L. B. et al. (2007). A new brief instrument for assessing decisional capacity for clinical research. *Archives of General Psychiatry*, *64*(*8*), 966–974. doi: 10.1001/archpsyc.64.8.966.

Johns, E. K., Phillips, N. A., Belleville, S., Goupil, D., Babins, L., Kelner, N. et al. (2012). The profile of executive functioning in amnestic mild cognitive impairment: Disproportionate deficits in inhibitory control. *Journal of the International Neuropsychological Society*, *18*(*3*), 541–555.

Kramer, J. H., Nelson, A., Johnson, J. K., Yaffe, K., Glenn, S., Rosen, H. J. et al. (2006). Multiple cognitive deficits in amnestic mild cognitive impairment. *Dementia and Geriatric Cognitive Disorders*, *22*(*4*), 306–311.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Nueropsychological assessment* (5th ed.). New York: Oxford University Press.

Maki, P., Rubin, L., Springer, G., Seaberg, E., Sacktor, N., Miller, E., . . . Study, Neuropsychology Working Groups of the Women's Interagency HIV Study and the Multicenter AIDS Cohort Study (2018). Differences in cognitive function between women and men with HIV. *Journal of Acquired Immune Deficiency Syndrome*, *79*(*1*), 101–107.

McCabe, D., Roediger, H., McDaniel, M., Balota, D., & Hambrick, D. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*, *24*(*2*), 222–243.

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, *32*, 541–554.

Moore, R. C., Kaufmann, C. N., Rooney, A. S., Moore, D. J., Eyler, L. T., Granholm, E. et al. (2016). Feasibility and acceptability of ecological momentary assessment of daily functioning among older adults with HIV. *American Journal of Geriatric Psychiatry*, *25*(*8*), 829–840. doi: 10.1016/j.jagp.2016.11.019.

Moore, R. C., Swendsen, J., & Depp, C. A. (2017). Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. *International Journal for Methods in Psychiatric Research*, *26*(*4*). doi: 10.1002/mpr.1562.

Paolillo, E. W., Obermeit, L. C., Tang, B., Depp, C. A., Vaida, F., Moore, D. J. et al. (2018a). Smartphone-based ecological momentary assessment (EMA) of alcohol and cannabis use in older adults with and without HIV infection. *Addictive Behavaviors*, *83*, 102–108. doi: 10.1016/j.addbeh.2017.10.016.

Paolillo, E. W., Tang, B., Depp, C. A., Rooney, A. S., Vaida, F., Kaufmann, C. N. et al. (2018b). Temporal associations between social activity and mood, fatigue, and pain in older adults with HIV: An ecological momentary assessment study. *JMIR Mental Health*, *5*(*2*), e38. doi: 10.2196/mental.9802.

Pilli, R., Naidu, M., Pingali, U. R., Shobha, J. C., & Reddy, A. P. (2013). A computerized Stroop test for the evaluation of psychotropic drugs in healthy participants. *Indian Journal of Psychological Medicine*, *35*(*2*), 180–189.

Repovš, G. (2004). The mode of response and the Stroop effect: A reaction time analysis. *Horizons of Psychology*, *13*(*2*), 105–114.

Schweitzer, P., Husky, M., Allard, M., Amieva, H., Peres, K., Foubert-Samier, A. et al. (2016). Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. *International Journal of Methods in Psychiatric Research*, *26*(*3*). doi: 10.1002/mpr.1521.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Reviews in Clinical Psychology*, *4*, 1–32.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(*2*), 461–479.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: administration, norms, and commentary*. American Chemical Society.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Reviews in Clinical Psychology*, *9*, 151–176. doi: 10.1146/annurev–clinpsy-050212-185510.

Uttl, B., & Graf, P. (1997). Color-Word Stroop test performance across the adult lifespan. *Journal of Clinincal and Experimental Neuropsychology*, *19*(*3*), 405–420.

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide Range Achievement Test-4: Professional Manual*. Lutz. FL: Psychological Assessment Resources, Inc.

Woods, S. P., Moore, D. J., Weber, E., & Grant, I. (2009). Cognitive neuropsychology of HIV-associated neurocognitive disorders. *Neuropsychology Review*, *19*(*2*), 152–168.