Comparison of PC and iPad administrations of the Cogstate Brief Battery in the Mayo Clinic Study of Aging: Assessing Cross-Modality Equivalence of Computerized Neuropsychological Tests

Nikki H. Stricker[1], **Emily S. Lundt**[2], **Kelly K. Edwards**[2], **Mary M. Machulda**[1], **Walter K. Kremers**[2], **Rosebud O. Roberts**[3,4], **David S. Knopman**[4], **Ronald C. Petersen**[4], and **Michelle M. Mielke**[3,4]

[1]Division of Neurocognitive Disorders, Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, USA

[2]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

[3]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA

[4]Department of Neurology, Mayo Clinic, Rochester, Minnesota, USA

Abstract

Objective: Computerized neuropsychological assessments are increasingly used in clinical practice, population studies of cognitive aging and clinical trial enrichment. Subtle, but significant, performance differences have been demonstrated across different modes of test administration and require further investigation.

Method: Participants included cognitively unimpaired adults aged 50 and older from the Mayo Clinic Study of Aging who completed the Cogstate Brief Battery and Cogstate's Groton Maze Learning Test (GMLT) on an iPad or a personal computer (PC) in the clinic. Mode of

Corresponding Author: Nikki H. Stricker, Ph.D., Mayo Clinic, 200 First Street SW, Rochester, MN 55905; 507-284-2649 (phone), 507-284-4158 (fax), Stricker.Nikki@mayo.edu.
Author Contact Details:
Nikki H. Stricker Division of Neurocognitive Disorders, Department of Psychiatry and Psychology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905; 507-284-2649 (phone), 507-284-4158 (fax), Stricker.Nikki@mayo.edu (email).
Emily S. Lundt Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Lundt.Emily@mayo.edu
Kelly K. Edwards Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, kellykedwards215@gmail.com
Mary M. Machulda Division of Neurocognitive Disorders, Department of Psychiatry and Psychology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Machulda.Mary@mayo.edu
Walter K. Kremers Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Kremers.Walter@mayo.edu
Rosebud O. Roberts Department of Neurology & Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Roberts.Rosebud@mayo.edu
David S. Knopman Department of Neurology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, knopman@mayo.edu
Ronald C. Petersen Department of Neurology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, peter8@mayo.edu
Michelle M. Mielke Department of Neurology & Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Mielke.Michelle@mayo.edu

administration differences and test-retest reliability coefficients were examined across 3 cohorts: a demographically-matched test-retest cohort completing PC and iPad administrations the same day (N = 168); a test naïve cohort comparing baseline PC (n = 1820) and iPad (n = 605) performance; and a demographically-matched longitudinal cohort completing 3 Cogstate visits over 15 months on either the PC (n = 63) or iPad (n = 63).

Results: Results showed a small but statistically significant and consistent finding for faster performance on PC relative to iPad for several Cogstate Brief Battery measures. Measures of accuracy generally did not differ or differences were very small. The GMLT showed faster performance and higher total errors on iPad. Most Cogstate variables showed no difference in the rate of change across PC and iPad administrations.

Conclusions: There are small, but significant, differences in performance when giving the same cognitive tests on a PC or an iPad. Future studies are needed to better understand if these small differences impact the clinical interpretation of results and research outcomes.

Keywords

Introduction

The use of computerized tests for cognitive screening and as part of neuropsychological evaluations is rapidly expanding (Crook, Kay, & Larrabee, 2009; Rabin et al., 2014; Snyder et al., 2011). Computerized tests offer several potential advantages over standard paper and pencil assessments. These advantages include greater efficiency, real time data entry, automated data exporting, automation of equivalent forms, the potential to reduce costs for test administration and scoring, easier adaptation to different languages, and suitability for at-home or long-distance use (Bauer et al., 2012; Fredrickson et al., 2010; Snyder et al., 2011; Wild, Howieson, Webbe, Seelye, & Kaye, 2008; Zygouris & Tsolaki, 2015). Current and potential applications of computerized neuropsychological assessment devices and the important limitations of their use have been reviewed (Bauer et al., 2012).

The same standards for traditional paper and pencil neuropsychological tests apply to the development, validation and use of computerized measures (Joint Committee on Standards for Educational and Psychological Testing, 2014; American Psychological Association, 2017). For example, like traditional tests, computerized measures should have demonstrated validity, adequate reliability, clear test administration instructions, interpretation guidelines, descriptions of normative or standardization samples, and adequate supporting documentation, e.g., test manual or technical report (Joint Committee on Standards for Educational and Psychological Testing, 2014).

The use of computerized tests also poses unique challenges. Because of the multitude of devices available for public use, multiple modes of administration are often available for the same computerized measure and it is unclear how well results generalize across different devices, operating systems, and other technical factors. Understanding subtle differences that may exist across testing platforms is important for clinical interpretation of test results and

for existing datasets that may contain cross-platform administrations of computerized neuropsychological tests. In addition, because many computerized tests target detection of longitudinal change, understanding whether longitudinal cognitive change differs by mode of administration is also critical.

There are several commercially available computerized testing platforms. Cogstate is one platform that focuses on the computerized measurement of cognition and is actively being used in clinical trials, epidemiological studies and clinics. Cogstate's platform offers a number of different tests and test batteries, including the Groton Maze Learning Test (GMLT) and the Cogstate Brief Battery (CBB) that are the focus of the present study. The CBB consists of four subtests that use a card game format to measure processing speed, attention, visual learning and working memory. The CBB recently received FDA approval under the name Cognigram™ as a Class II Exempt Medical Device intended to serve as a digital cognitive assessment tool for prescription use with self-administered assessment in individuals aged 6–99. Marketing materials available online (www.Cogstate.com) encourage its use for both single assessments and change over time, with support for in-clinic or at-home testing. The CBB is identical to the Axon Sports/Cogstate Sport battery used within the sport-related concussion literature (Nelson et al., 2016). Prior studies have raised concerns about low reliability coefficients for CBB measures in samples of adolescents and young adults (Cole et al., 2013; Fratti, Bowden, & Cook, 2017; MacDonald & Duerson, 2015; Nelson et al., 2016), although high to very high reliability values were reported in one sample of young adults with a 1 week test-retest interval (Louey et al., 2014) as well as in older adults (Fredrickson et al., 2010; Y. Y. Lim et al., 2013). It is presumed that these prior studies were completed on PCs, although this is frequently not explicitly stated within studies. Comparison of reliability of Cogstate subtests on PC versus iPad is therefore needed.

Cogstate can be administered using a variety of modalities (i.e., PC, Mac, tablet device) and web browsers. However, limited data are available regarding mode of administration differences for Cogstate or for computerized neuropsychological tests more broadly. The NIH Toolbox-Cognition recently switched from a web-based version administered on a PC to iPad administration. Using a stratified random assignment equivalent-groups design, some significant differences across platforms were reported in preliminary analyses (Gershon, Nowinski, Kaat, & Slotkin, 2017) and several adjustments to normative data were subsequently applied. Specifically, better scores were seen on the iPad on the Flanker Inhibitory Control and Attention Test and the Dimensional Change Card Sort Test, and worse scores were seen on the iPad for the Pattern Comparison Processing Speed Test and the Picture Sequence Memory Test. To our knowledge, our group has published the only available data to date that investigates potential differences across different modes of administration using Cogstate (Mielke et al., 2015). We showed small but statistically significant performance differences on the PC compared with the iPad on several Cogstate tasks. Currently, the normative data available from Cogstate do not consider mode of administration, which could pose problems for clinical application and interpretation.

The Mayo Clinic Study of Aging began implementing Cogstate measures in 2012, specifically the CBB and the GMLT. We have demonstrated that Cogstate is both feasible

and valid within the Mayo Clinic Study of Aging (Mielke et al., 2015), and that performance on Cogstate is related to measures of neurodegeneration (Mielke et al., 2014). We initially administered Cogstate on a PC. However, based on participant feedback that demonstrated a clear preference for iPad administration relative to PC and that the iPad is easier to use for those with arthritis (Mielke et al., 2015), we switched to the iPad for all in-clinic administrations of Cogstate by 2015. We previously presented preliminary data comparing same day test-retest administration of Cogstate across these two different platforms, showing a few small, but statistically significant differences (Mielke et al., 2015). A more in-depth look at these differences is the focus of the current manuscript. This study has several aims. First, we expanded upon our prior comparisons of same day, within-subject test-retest Cogstate comparisons across PC and iPad administrations with updated inclusion criteria to minimize potential confounds. All participants within this cohort had at least one prior exposure to Cogstate, thus this comparison allowed a direct look at cross-modality differences on the same day while practice effects were minimized as participants were already familiar with the test. Second, we aimed to determine whether a similar pattern of results is seen in test naïve participants by using a cross-sectional design that takes advantage of all our available data and allowed us to focus on potential cross-modality differences upon initial exposure to Cogstate, which may amplify any potential human factor influences as individuals become familiar with the task. Third, we investigated whether there are any differences in change over time across modes of administration.

Method

Participants

The Mayo Clinic Study of Aging (MCSA) is a population-based study of cognitive aging among Olmsted County, MN, residents that began in October 2004, and initially enrolled individuals aged 70 to 89 years with follow-up visits every 15 months. The details of the study design and sampling procedures have been previously published; enrollment follows an age- and sex-stratified random sampling design to ensure that men and women are equally represented in each 10-year age strata (Roberts et al., 2008). In 2012, enrollment was extended to cover the ages of 50–90+ following the same sampling methods. The MCSA began administering Cogstate measures in 2012 to the newly enrolled 50–69 year olds on a PC during clinic visits. Cogstate administration for those aged 70 and older began the following year. In 2014 we piloted administration of Cogstate on an iPad by administering Cogstate on both a PC and an iPad during the same study visit (see test-retest sample below) among MCSA participants. This piloting yielded participant feedback that demonstrated a clear preference for iPad administration of the Cogstate relative to the PC (Mielke et al., 2015). It was also apparent that the iPad was easier to use for those with arthritis, particularly for the GMLT. Because of this feedback, we began to switch to iPad administration for all Cogstate tasks in 2014, and this transition was complete by early 2015. The current study included cognitively unimpaired individuals aged 50 or older who completed the CBB in the clinic on a PC or iPad. Individuals under 75 also completed Cogstate's GMLT. Initial piloting of the GMLT in individuals over 75 revealed that vision and arthritis (e.g., difficulty manipulating the mouse) were frequent confounds for the GMLT in this age group. Baseline Cogstate was typically completed at a Mayo Clinic Study

of Aging study visit. For longitudinal follow-up, Cogstate is also administered in the clinic in between Mayo Clinic Study of Aging study visits, approximately 7.5 months later. To summarize, at baseline and all follow-up visits, individuals under 75 completed both the CBB and the GMLT at each visit. Individuals 75 and older completed only the CBB at baseline and all follow-up visits.

The study protocols were approved by the Mayo Clinic and Olmsted Medical Center Institutional Review Boards. All participants provided written informed consent.

Study visits included a neurologic evaluation by a physician, an interview by a study coordinator, and neuropsychological testing by a psychometrist (Roberts et al., 2008). The physician examination included a medical history review, complete neurological examination, and the administration of the Short Test of Mental Status (Kokmen, Smith, Petersen, Tangalos, & Ivnik, 1991). The study coordinator interview included demographic information and medical history, and questions about memory to both the participant and informant using the Clinical Dementia Rating scale (Morris, 1993). See Roberts et al. (2008) for details about the neuropsychological battery.

For each participant, performance in a cognitive domain was compared with age-adjusted scores of cognitively unimpaired individuals using Mayo's Older American Normative Studies (Ivnik et al., 1992). Participants with scores of 1.0 SD below the age-specific mean in the general population were considered for possible cognitive impairment, taking into account education, prior occupation, visual or hearing deficits, and other information. Diagnosis of mild cognitive impairment (MCI) or dementia was based on a consensus agreement between the interviewing study coordinator, examining physician, and neuropsychologist, after a review of all participant information (Petersen et al., 2010; Roberts et al., 2008). Individuals who performed in the normal range and did not meet criteria for MCI or dementia were deemed cognitively unimpaired and were eligible for inclusion in the current study. Performance on the computerized cognitive battery was not considered for diagnosis.

Cogstate Computerized Battery

Cogstate has several computerized tests available for use. This study used the CBB that includes four card tasks (see below) and the GMLT (Fredrickson et al., 2010; Lim et al., 2012; Pietrzak et al., 2008; Snyder, Bednar, Cromer, & Maruff, 2005). Given previous literature showing an initial practice/learning effect between the first and second administration (Collie, Maruff, Darby, & McStephen, 2003), we administered a short practice battery, followed by a 2-minute rest period, then the complete battery. A short practice battery is included each time Cogstate is administered and is not used in any analyses. The study coordinator was available to help the participants understand the tasks during the practice session. During the test battery, the coordinator provided minimal supervision or assistance.

The tests were administered in the order listed below. The data were automatically transformed by Cogstate for the four card tasks in the CBB using logarithmic base 10 transformation for reaction time data (collected in milliseconds) and arcsine transformation

for accuracy data to normalize the variables (see Collie et al., 2003; Fredrickson et al., 2010; Lim et al., 2012; Pietrzak et al., 2008 for additional description). The data were not age adjusted. Cogstate recommends a primary outcome variable for each measure. There have been some changes over time regarding which variable for any subtest is viewed as the primary outcome measure. Due to the exploratory nature of these analyses as well as the lack of consistency in the literature a comprehensive approach to variable inclusion was preferable. Measures of interest included recommended primary outcomes as well as secondary outcome variables as described below. Untransformed (raw) scores were also included to increase interpretability of platform differences. Response time variability (i.e., the standard deviation of speed of performance) is presented for transformed speed variables.

Cogstate Brief Battery—Detection is a simple reaction time paradigm that measures psychomotor speed. Participants press "yes" as quickly as possible when a playing card turns face up. Reaction time (i.e., speed) for correct responses was the primary outcome measure. Accuracy was a secondary outcome measure.

Identification is a choice reaction time paradigm that measures visual attention. Participants press "yes" or "no" to indicate whether or not a playing card is red as quickly as possible. Reaction time for correct responses was the primary outcome measure. Accuracy was a secondary outcome measure.

One Card Learning is a continuous visual recognition learning task that assesses learning and attention. Participants press "yes" or "no" to indicate whether or not they have seen the card presented previously in the deck. Accuracy was the primary outcome measure. Reaction time for correct responses was a secondary outcome measure. Task instructions do not indicate a need to respond as quickly as possible.

One Back assesses working memory. Participants press "yes" or "no" to indicate whether or not the card is the same as the last card viewed (one back) as quickly as possible. Reaction time for correct responses was the primary outcome measure based on the recommendation of Cogstate. Accuracy was a secondary outcome measure in the current study and has been used as a primary outcome measure in other studies (Y. Y. Lim et al., 2013; Maruff et al., 2013).

Groton Maze Learning Test—The GMLT is a hidden pathway maze learning test that measures spatial working memory, learning efficiency, and error monitoring (Pietrzak et al., 2008; Snyder et al., 2005). The primary outcome measure was the average correct GMLT Moves Per Second (MPS; speed/efficiency) across the five trials, reflecting learning efficiency. Higher MPS indicate faster response time. The GMLT Total Errors (e.g., total number of errors) was a secondary outcome measure that reflects working memory and/or executive function and has been used as a primary outcome measure in other studies (Lim et al., 2015).

The Cogstate battery provides a large number of equivalent alternative forms by having a large stimulus set from which exemplars are randomly chosen at run time, resulting in a

different set of exemplars that are used each time an individual takes the test. The correct response (either yes or no) is randomly chosen for each trial of the task, and the interstimulus interval has a random interval that varies for each trial of the task. For the GMLT there are 20 possible hidden pathways matched for number of tiles and turns. These are presented in a random, nonrecurring order yielding 20 equivalent alternative forms.

The response input for Cogstate tasks varied by mode of administration. For the PC, a keyboard was used for the four CBB card tasks and a mouse was used for the GMLT. On the iPad, the input was by finger touch for the four CBB card tasks and by stylus for the GMLT.

The ability to reliably complete and adhere to the requirements of each task was determined by completion and integrity checks. For each card task in the CBB (Detection, Identification, One Card Learning, One Back) a completion criteria of 75% or more of the trials within the task was required. For the GMLT   140 correct moves was required (i.e., all five trials of the test had to be completed within 10 minutes). For integrity flags, per the test developers and previous publications, data on each specific test was determined valid if the accuracy of Detection was >90%, Identification was >80%, One Card Learning was >50%, and One Back was >70%. The GMLT data were considered valid if Total Errors were < 120 (Fredrickson et al., 2010; Mielke et al., 2015; Pietrzak et al., 2008).

Study Design and Statistical Methods

We compared performance on Cogstate administered on the iPad versus the PC using several approaches to fully investigate potential mode of administration differences. Analyses were conducted using statistical software R version 3.4.1.

Test-Retest Sample—We compared performance intra-individually via a test-retest alternate forms design. The CBB and GMLT were administered on a PC and iPad consecutively at the same study visit in the clinic, with a 2- to 3-minute break between. Previously, a portion of this data (n = 341) was presented as part of a description of Cogstate performance and its relationship to a standard neuropsychological battery within the Mayo Clinic Study of Aging (Mielke et al., 2015). Study protocol was to alternate the administration order. Despite this protocol, it was later discovered that 75% of these participants were tested on the PC first. To minimize the potential bias of unbalanced groups, we applied a matching procedure in the current study and implemented 1:1 matching for PC-first and iPad-first, number of prior Cogstate visits, age, education and sex. Following application of the matching procedure, the sample included 168 individuals, 84 of whom completed the PC administration first. All participants within this cohort completed Cogstate on a PC at least once prior to this same day test-retest visit, with prior test exposures ranging from 1 to 5 (See Table 1). The current study presents additional variables not examined previously.

To compare the frequency of completion failures and integrity failures across PC and iPad administrations, we computed Chi-square tests of independence. All data values with a failed completion or integrity flag were removed from subsequent analyses.

Differences were computed within subject for each outcome variable as iPad score minus PC score and subsequently analyzed. Student's t-tests were used to assess whether the computed differences statistically differed from zero indicating that the two tests were not comparable. Standardized mean difference Cohen's d and 95% confidence intervals (CI) of effect size were computed.

To determine whether any differences across modes of administration were related to demographic variables, tests for evidence of an effect modifier were calculated using Pearson's correlation coefficient rho for age and education. For sex, categorical, linear regression models were fit between computed differences and sex.

To assess cross-platform reliability, we examined Bland-Altman plots, Pearson's correlation coefficients (rho) and intraclass correlation coefficients (ICC).

Test Naïve Sample—To eliminate potential practice effects on observed platform differences we analyzed baseline Cogstate data. There were 1820 participants who completed the CBB on the PC and 605 on the iPad at baseline. PC and iPad groups differed on age and education, but were comparable on sex (see Table 2). There were 1634 participants who completed the GMLT on the PC and 195 on the iPad at baseline. For the GMLT, PC and iPad groups differed on age, but not education or sex. The PC group was younger because of the timing of when the PC versus the iPad was administered. The initiation of Cogstate administration coordinated with the expansion of the MCSA cohort to 50–69 year olds in 2012. We initially only administered the Cogstate tests to this cohort because we wanted to assess the feasibility, acceptability and response in this population-based community sample among middle-aged adults prior to trying to administer to those aged 70 and older with typically less computer experience. After we determined the feasibility and acceptability of completing Cogstate, we then began administrating to individuals aged 70 and older, about 6–12 months later. Chi-square tests of independence were computed to compare frequency of failure to meet completion and integrity criteria. All data points with a failed completion or integrity flag were removed from subsequent analyses.

Linear regression models were fit on each outcome measure. To account for demographic differences, age, sex, and education were included as covariates in the models. We tested for evidence of interactions (platform by age, platform by education, and platform by sex) but none were significant at the 0.05 level so were excluded from the final fitted model. Given a large dynamic range on age we included a 3-knot restricted cubic spline on age in order to allow for non-linearity in age.

Longitudinal Comparison—Participants completed the Cogstate battery on either an iPad or PC in the clinic at baseline and over two follow-up visits approximately 7.5 and 15 months later. Individuals completed all visits on the same platform (e.g., either all PC visits or all iPad visits). In order to minimize any potential influence of demographic variables, each participant tested on iPad was matched by age, sex, and education to one participant on PC resulting in 126 total participants, yielding groups that were comparable on age, education and sex. The iPad group had a slightly shorter overall duration of follow-up (see

Table 3). A smaller sample size was available for the GMLT, with 59 participants completing the GMLT on PC and 24 on iPad at baseline, and the PC group was slightly older than the iPad group. To assess whether iPad and PC outcomes were comparable at each time point we computed Student's T-tests for mean group difference and provide Cohen's d estimates of effect size. Results were the same using Wilcoxon-Rank Sum tests, thus we report only T-test results.

Since participants were tested at regular intervals (every 7.5 months) we used a reduction of variables approach to assess change over time. We fit linear regression models in each participant and captured the slope as a measure of annual change, repeating this for each outcome. Mean slopes for each outcome variable were tested for platform differences using Student's T-tests. To explore the possibility of nonlinear relationships with time, we also fit linear mixed effects models adjusting for baseline age, sex, education, platform of testing (PC or iPad), time, time-squared, and a 3-way interaction among platform, time and time-squared. Our conclusions remained the same after inspecting results of both modeling approaches, thus we only present the results of the linear regression analysis. The longitudinal dataset is comprised of subjects tested three times at 7.5 month intervals making it well balanced over time, thus mixed models and the reduction of variables method yield very similar results.

We report reliability as Pearson's correlation coefficient (rho) and intraclass correlation coefficients (ICC) for baseline to 7.5 months and 7.5 months to 15 months.

Results

Test-Retest Results

Completion and Integrity Flags—There were no differences in the frequency of failing completion or integrity flags across mode of administration (see Table 1).

Cogstate Brief Battery Mode of Administration Comparisons—Participants performed significantly faster on the PC than iPad on all CBB measures (Detection speed, Identification speed, One Back speed, One Card Learning speed), with Cohen's d effect sizes ranging from 0.19–0.41 (see Table 4). There were no differences in accuracy across mode of administration for any of the CBB card tasks. Variability in response speed did not differ for any CBB card task and raw score comparison results were equivalent to transformed variables.

GMLT Mode of Administration Comparisons—A different pattern of findings were seen on the GMLT; participants were faster (i.e., had higher Moves Per Second) and made more Total Errors on the iPad than the PC.

Correlations with Demographic Variables—Correlations were weak and not significant between scores across iPad minus PC performance and age or education (all r's < 0.14). Sex was marginally associated with GMLT Total Errors ($p = .06$), with males having on average 4 more errors (95% CI = –0.16–8.90) on iPad versus PC compared to females. No other sex effects were demonstrated.

Reliability—Test-retest reliability coefficients ranged from low to adequate (see Table 5). Inspection of Bland-Altman plots for continuous measures suggests the two platforms have similar distributions with mean difference centered at 0 and few values outside the 95th percentile confidence interval, with the exception of GMLT.

Test Naïve Results

Completion and Integrity Flags—See Table 2 for completion and integrity flag failure rates. Higher rates of completion and integrity failures were seen across nearly all measures in the iPad group and may be related to the older age and fewer years of education of that group given the lack of any such differences in our demographically-matched test-retest sample as reported above. Primary results described below control for these important age and education differences across the PC and iPad groups.

Cogstate Brief Battery Group Comparisons—See Table 6 for group difference estimates accounting for covariates. The PC group performed significantly faster than the iPad group across several CBB measures, including Identification speed, One Back speed, and One Card Learning speed. The groups did not differ on Detection speed or One Card Learning accuracy. The iPad group had higher accuracy on Identification and One Back, whereas the PC group had higher accuracy on Detection.

GMLT Group Comparisons—On the GMLT, the iPad group was faster (i.e., higher Moves Per Second) and made more Total Errors relative to the PC group.

Longitudinal Results

Completion and Integrity Flags—There were no significant differences in completion or integrity failure rates between PC and iPad groups across all visits (see Table 3).

Cogstate Brief Battery Group Comparisons—Most participants had data from all 3 time points available for the CBB (81–94% across measures; see Table 7). There was no difference in rate of change over time across modes of administration for most primary Cogstate outcome variables (Detection speed $p = .90$, Identification speed $p = .09$, One Card Learning accuracy $p = .31$), but a difference emerged for one primary variable (One Back speed $p = .001$). For One Back speed, participants performed more slowly on the iPad at baseline but then appeared to "catch up" to the PC group at subsequent visits, with no differences apparent at time 2 or time 3. All secondary outcome variables showed no difference in rate of change over time across PC and iPad administrations (Detection accuracy $p = .69$, Identification accuracy $p = .69$, One Card Learning speed $p = .34$, One Back accuracy $p = .32$).

Group differences across PC and iPad groups at each time point are denoted in Figure 1 and means, SDs, p-values and effect sizes are available in Table 7 for all variables. At baseline, the PC group performed faster than the iPad group across most CBB card tasks (Identification speed, One Back speed, One Card Learning speed; medium to large effect sizes), except for Detection speed. There were no speed differences on CBB card tasks across PC and iPad groups at time 2 or time 3, with the exception of faster Identification

speed on the PC at time 3 (medium effect size). There were generally no differences in accuracy for any CBB card tasks (except Detection accuracy at Time 1).

GMLT Group Comparisons—Only 16 participants had all 3 GMLT data points available (also see Table 7 for n's by mode of administration for each time point). These data are presented, but significant caution is needed given the small sample size and restricted age range. There was no difference in rate of change over time across modes of administration for GMLT Moves Per Second ($p = .29$) or GMLT Total Errors ($p = .09$). However, visual inspection of GMLT Moves Per Second suggests there may be a trend that could be captured in future studies with a larger sample. The iPad group was faster (higher Moves Per Second) at baseline and time 2 than the PC group; this group difference was not significant at time 3 although this may have been due to the small sample size at that time point. GMLT Total Errors did not show group differences at any time point.

Reliability—Test-retest reliability coefficients ranged from low to high across subtest and platform (see Table 5). Reliability coefficients for speed on card tasks on the PC were higher, typically in the marginal to adequate range (although Identification values were in the low range) relative to the iPad, which were generally in the low to marginal range. Reliability values for One Card Learning accuracy were low, and somewhat higher on the iPad than PC. Reliability values for One Back accuracy are reported but should be interpreted cautiously due to the non-normal distribution of this data. GMLT reliability values are generally high for baseline to time 2, with slightly lower values on iPad for Total Errors. Time 2 to time 3 reliability coefficients are reported but should be interpreted cautiously for the GMLT due to small sample sizes.

Discussion

The main finding of this study is that there are small, but significant, differences in performance on Cogstate measures when administered on a PC versus an iPad, primarily for timed outcome measures. In accordance with our study aims, this was (1) demonstrated within a test-retest sample, (2) generally replicated in a test naïve cohort, and (3) longitudinal analyses showed largely stable mode of administration differences over time for most primary outcome variables.

Repeat administration of Cogstate on a PC and iPad on the same day showed that the same participants performed significantly faster on the PC than the iPad across all CBB card tasks. Effect sizes ranged from minimal to small. This pattern of results was mostly replicated across the test naïve and longitudinal cohorts, but some subtle and possibly important differences emerged. Specifically, there were no differences in Detection speed, a simple reaction time paradigm, at baseline or follow-up visits within the longitudinal sample. At baseline assessment, as with the test-retest sample, differences were seen on all other CBB card tasks for speed outcome variables. Interestingly, within our longitudinal sample these differences largely were absent across follow-up visits. The slight variability in our results depending on the cohort and time point suggests that human factors, as opposed to hardware differences, may drive some of these differences. In support of this possibility, Burke et al. (2017) also reported faster choice reaction time responses on a traditional visual choice

reaction time apparatus device versus an iPad/iTouch, but no differences across these modes of administration on a simple reaction time task. Valdes, Sadeq, Harrison Bush, Morgan, & Andel (2016) trained older adults to self-administer the CBB at home monthly for one year and found greater intraindividual variability at baseline, particularly for Detection and Identification speed. When the baseline assessment was excluded, intraindividual variability was stable across sessions. We similarly found a significant and large (d = −.85) difference in Detection response time variability at baseline in the longitudinal sample, with greater variability on the iPad than on the PC, which was not present at follow-up visits. This suggests that performance may be more impacted by different human factor variables at baseline when individuals are continuing to acclimate to the test and device. The absence of any differences in variability within the test-retest sample (who had at least one prior exposure to Cogstate) further suggests than human factor variables / individual differences in approaching the different response input requirements across devices is likely greatest upon initial exposure to the tasks; the short practice trial may not be enough to fully attenuate these differences. For example, a touch response is required on the iPad for all CBB card tasks, and Detection is the first task administered. There are no instructions or standards for hand placement during these tasks, thus the variability across modes of administration upon first exposure to the CBB card tasks may be largest during the Detection task, and then variability decreases as participants acclimate to a presumably more consistent hand position and response pattern.

These baseline differences may also be related to the degree of prior exposure to computers/ technology, which may be less in older individuals. Related to this, Valdes et al. (2016) also found that individuals most comfortable with computers were faster responders on Detection than those less comfortable with computers, and that more frequent computer users were more accurate on the One Back task and showed less of a practice effect on both One Back and One Card Learning accuracy over time. It is possible that the standard practice session administered by Cogstate may not be enough to mitigate some of these individual differences and further consideration should be given to administration of an additional baseline assessment that is not used for primary analyses. Within the current study, we did not see evidence of any interaction with age and PC-iPad differences, but future studies focused on potential mode of administration differences would benefit from including a measure of frequency of computer/technology use and comfort with each device participants are tested on. Inclusion of a wider age range of participants would also be helpful to determine whether mode of administration differences are attenuated in younger adults who typically have higher levels of computer/technology use and familiarity.

It is possible that at least some of the difference in speed observed on the CBB card tasks could be explained by hardware and/or response input differences, as participants used the keyboard to respond on PC CBB card tasks versus touch on iPad. This assumption requires verification with automated timing devices (e.g., external chronometers) or data visualization methods that can help identify inaccuracies in timing (Salmon et al., 2017). This is particularly relevant for reaction time measures, such as Detection and Identification speed. See Cernich, Brennana, Barker, & Bleiberg (2007) and Parsons, McMahan, & Kane (2018) for a detailed review of how different software and hardware configurations and input devices can impact computerized neuropsychological assessment and for discussion of

several steps that can be taken to help mitigate this error. Despite the clear possibility that these device-related factors may drive these findings, there were no differences in accuracy across modes of administration for any of the CBB card tasks within the test-retest sample. In the longitudinal sample, accuracy differences on CBB card tasks were also largely absent, with the exception of Detection accuracy at baseline. Within the test naïve sample, no differences were seen on One Card Learning accuracy, but Detection, Identification and One Back accuracy showed significant differences. However, the magnitude of this difference was quite small, ranging from 0.3 – 0.6% based on the raw accuracy estimates across platforms (see Table 6), and it is likely that these significant results are due to our large sample sizes within this cohort. In addition, Detection and Identification accuracy variables are not primary outcome variables, and are rarely reported in other Cogstate studies. Overall, we interpret our data as suggesting there are minimal differences in accuracy across modes of administration. Thus, accuracy data can be readily combined across participants tested on either a PC or iPad, without need for any additional corrections.

Despite these minimal accuracy differences on CBB card tasks, it cannot be assumed that all untimed tasks will show equivalent performances across different modes of administration as a different pattern emerged for the GMLT. Within both the test-retest and test naïve samples, there were a higher number of errors on the iPad relative to the PC, although the effect size was small. In addition, speed differences were generally reversed on this task, with faster performance on the iPad (higher MPS) relative to the PC and a large effect size for this difference in the test-retest sample. The response input on this task is different from the CBB card tasks, with a stylus for the iPad and a mouse for the PC. It may be that use of the stylus results in a faster but more impulsive response style.

Age, education and sex were not associated with any performance differences across the PC and iPad within the test-retest sample, although there was a trend showing that males made more errors on the iPad versus PC relative to females on the GMLT. Within the test naïve sample, there were no significant interactions of these demographic variables and mode of administration. However, all Cogstate variables were strongly associated with age, consistent with prior findings (Dingwall, Lewis, Maruff, & Cairney, 2010; Fredrickson et al., 2010). In addition, most Cogstate outcome variables were also significantly associated with education and a few were associated with sex. Like other studies showing this association (Cromer, Schembri, Harel, & Maruff, 2015), the size of this relationship was small and the significance of the association is likely related to our large sample size and high power to find a small relationship with these demographic variables within the test naïve sample (e.g., it was typically but not always smaller in magnitude than the platform difference).

To our knowledge, no prior studies have reported reliability coefficients across different modes of administration. Test-retest reliability coefficients within our longitudinal sample ranged from low to high across subtests and platform. Reliability coefficients tended to be higher for repeat PC administrations relative to repeat iPad administrations. Values for Pearson's r and ICC were similar. Our results are also similar to prior studies presenting data for younger populations reporting that several reliability coefficients fall below levels required for clinical decision making (Cole et al., 2013; Fratti et al., 2017; MacDonald & Duerson, 2015; Nelson et al., 2016). This was particularly the case for measures of accuracy

(One Card Learning, One Back), as also observed in other studies (Fratti et al., 2017). Given that these accuracy values are also most likely to show practice effects relative to other Cogstate variables (Fredrickson et al., 2010; Mielke et al., 2016), it may be that practice effects contribute to lower reliability values (Slick, 2006). One Back accuracy is also non-normally distributed, which may artificially lower reliability (Slick, 2006). In addition, because Cogstate uses alternate forms in an effort to reduce practice effects this may attenuate reliability coefficients to some extent (Calamia, Markon, & Tranel, 2013). Reliability values for consecutive testing within the same day across different modes of administration were also low to adequate, despite the limited time between assessments. This suggests that the small performance differences across modes of administration may have a negative impact on reliability values, although our data does not allow a direct comparison of this assumption given the different time intervals of within versus across mode of administration reliability coefficients reported in this manuscript. Overall, reliability values reported here are generally lower than those provided in other studies of Cogstate performance in older adults, although this may be because those studies report the average ICC across several follow-up intervals (Fredrickson et al., 2010; Lim et al., 2013). Continued investigation of the reliability of Cogstate and other computerized measures both within and across different modes of administration is needed. Lower reliability within platform, as demonstrated here for iPad administration, will further weaken reliability when individuals perform assessments across different modes of administration over time.

It is important to determine whether the small mode of administration differences reported here represent clinically meaningful differences, or simply add a small amount of noise to the test data – both for the clinical interpretation of individual test results and for assessing research outcomes. Further studies would be useful in this regard but our data provide some insight into this issue and we provide effect sizes to help determine the magnitude of these differences, as discussed above. Another way to examine the magnitude of these differences is to see how the average mode of administration difference may impact normative scores and norms for change. Cogstate provides normative data in the form of means and standard deviations to users, and the latest iteration of this normative data is based on data collected through November, 2017 (Cogstate, 2018). Cogstate reports that the adult normative sample for individuals aged 18 to 99 years is based on a healthy population of participants enrolled in a series of clinical trial, research and academic studies from countries in North and South America, Europe, Asia and Australia. The data collapse across different modes of administration but predominantly represent administration on PC. Means and standard deviations by different age bands are provided for baseline assessments, and all participants completed at least one practice assessment prior to their baseline assessment. Using the mean PC-iPad difference from our test-retest sample as reported in Table 4 and the normative data provided by Cogstate for individuals aged 60–69 (all n's > 4500 for each card task, n = 329 for GMLT errors), the mode of administration difference alone would result in a z-score difference for speed of 0.24 on Detection, 0.31 on Identification, and 0.27 on One Back. The z-score differences for accuracy on CBB card tasks are negligible, with 0.07 for One Card Learning accuracy and 0.01 for One Back accuracy. The z-score difference for GMLT Total Errors is 0.14; normative data are not provided for GMLT Moves Per Second. The Cogstate normative data also provide within-subject standard deviation

(WSD) values for most primary outcome variables in the normative data. This can be used in a reliable change index formula to calculate a z-score for change (Hinton-Bayre, 2010; Lewis, Maruff, Silbert, Evered, & Scott, 2007). The approximate test-retest interval is 1 month, thus some caution is needed in the interpretation of the following values that are based on same day administration. Using the mean PC-iPad difference from our test-retest sample and the WSD values provided, the mode of administration difference alone yielded z-scores for change of 0.26 on Detection speed, 0.44 on Identification speed, and 0.37 on One Back speed. One Card Learning Accuracy yielded a z-score for change of 0.07, and data were not available for One Back Accuracy. GMLT Total Errors yielded a z-score for change of 0.14. These examples are provided to help illustrate the magnitude of mode of administration differences. The degree to which this is clinically relevant may depend on factors such as where the individual's performance falls in the normative distribution, what the score is being used for, and whether or not there are changes in mode of administration over time within the same individual.

We also offer some preliminary recommendations for addressing the small platform differences demonstrated in the current study that we will apply to the MCSA. First, not all Cogstate measures show consistent differences across mode of administration. Thus it seems unnecessary to require generation of separate normative data and norms for change for each platform. We plan to apply a linear correction for variables showing consistent platform differences based on our test-retest cohort's mean differences across platforms to each PC time point to allow us to collapse across all data. Specifically, we will apply a correction for Identification speed, GMLT Moves Per Second, and GMLT Total Errors. Differences in accuracy on CBB card tasks are negligible, thus no correction is needed for One Card Learning accuracy and One Back accuracy. Because of this, we will use One Back accuracy as our primary outcome variable for the One Back task, as opposed to One Back speed, which showed consistent platform differences. Similarly, preference will be given to GMLT Total Errors as the primary outcome variable for GMLT since the magnitude of platform differences was smaller for that variable. We do not plan to use a correction for Detection speed; although the test-retest sample did show a small difference, the test naïve and longitudinal analyses showed no significant differences for this measure and we prefer to take the most conservative approach to changes to the raw data possible. Based on our results, we recommend that clinical trials use only one mode of administration throughout the study to maximize reliability and reduce device-related error. Consideration may also be given to maintaining consistency within that mode of administration by keeping the operating system and input devices the same across the clinical trial, where feasible.

Strengths of our study include the large sample size, use of different study designs to investigate potential mode of administration differences, and the population-based nature of our study that facilitates generalization of these findings. However, limitations should also be noted. There may be subtle effects of historical cohort differences on the current results because PC data tended to be collected earlier (prior to 2015) relative to iPad data. This historical difference also restricted the age ranges available for some analyses given differences in the age of participants emphasized for recruitment over time. This age difference was evident in our test naïve sample, where a nearly 10-year age difference was present between baseline PC and iPad groups. We also saw education differences in the test

naïve sample. Although we corrected for these age and education differences in our models examining PC vs. iPad group differences in the test naïve sample, a prospective approach ensuring matched groups on important demographic variables would have been more ideal. However, we used a 1:1 matching procedure to ensure matched demographic variables within our test-retest and longitudinal samples. Thus, although we cannot ensure that covarying age and education did not impact the test naive results to some degree, comparison of these results with our demographically matched test-retest and longitudinal samples suggests this is unlikely.

This was a retrospective study that took advantage of available data to examine differences across mode of administrations. Because of the retrospective nature of the study, combined with the historical cohort differences described above, we unfortunately do not have the data necessary to fully parse apart potentially important hardware, software and device-input related differences that may contribute to these results beyond the broad PC vs. iPad comparison that includes the different device inputs for each (Cernich et al., 2007; Parsons et al., 2018). For example, although there are external chronometers and timing analysis tools (e.g., Black Box Toolkit) that are available to assist with determining the amount of reaction time variability present (Parsons et al., 2018) and that would have been helpful to use across PC and iPad devices, applying these measures retroactively is not feasible as operating systems have been updated and multiple different PC computers and iPads were used during this study without tracking which device was used for each participant and time point. Methods of visualizing reaction time data can also assist with detecting temporal inaccuracy across different hardware/software systems (Salmon et al., 2017). However, this would also require knowing specific device information for each administration, as well as access to individual trial by trial results that are not part of the standard Cogstate output.

It would be beneficial for future studies to examine split half reliability and within-session practice effects for possible differences across modes of administration. Collie et al. (2003) demonstrated that simple reaction times became significantly faster during the first assessment, and then leveled out at retest on a measure similar to Detection speed. The inclusion of an initial practice trial that is excluded from analysis, as is standard practice for Cogstate measures, helps to mitigate this type of practice effect. It remains unclear whether this type of within-session practice effect occurs again for the first exposure to a different test modality.

The generalizability of our results is limited in a few important ways. First, our sample is predominantly White and not of Hispanic origin (99% in the test-retest sample; 97% in the test naïve and longitudinal samples). Our sample also was limited to cognitively unimpaired individuals. Future studies should investigate whether platform differences are similar or amplified in clinical populations, such as in individuals with mild cognitive impairment and Alzheimer's dementia, and should replicate these findings in more diverse samples. Finally, we did not administer the GMLT to individuals over 75. Therefore, available sample sizes for that measure were smaller, particularly for longitudinal analyses. Those results should be viewed as preliminary and require verification with a larger sample.

In summary, the current study describes slight, yet significant mode of administration differences for computerized neuropsychological tests. These differences will be important for test developers and users to further assess, understand, and adjust for when needed. The current data suggest that tracking and reporting what type of device is used may be a critical step to ensure that mode of administration differences can be fully understood and accounted for when developing and using computerized neuropsychological tests and when interpreting research results. We suggest that test developers take advantage of available methods for automatically detecting and recording the operating system and version being used by end users whenever possible. Test developers are also encouraged to examine their software's performance across different devices (e.g., including different testing platforms, monitors, input devices, etc.), document results in test manuals or technical documents, and provide clear minimum system requirements for users (Parsons et al., 2018). Just as direct translation of a paper and pencil neuropsychological measure to computerized administration requires a demonstration of test equivalency (Bauer et al., 2012), our results suggest that adaptation of computerized measures to different platforms or response inputs similarly requires empirical demonstration of comparability of the outcome data.

Acknowledgements

References

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct (2002, Amended June 1, 2010 and January 1, 2017)

Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, & Naugle RI (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. The Clinical Neuropsychologist, 26(2), 177–196. doi: 10.1080/13854046.2012.663001 [PubMed: 22394228]

Burke D, Linder S, Hirsch J, Dey T, Kana D, Ringenbach S, et al. (2017). Characterizing information processing with a mobile device: measurement of simple and choice reaction time. Assessment, 24(7), 885–895. doi: 10.1177/1073191116633752 [PubMed: 26933140]

Calamia M, Markon K, & Tranel D (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. The Clinical Neuropsychologist, 27(7), 1077–1105. doi: 10.1080/13854046.2013.809795 [PubMed: 24016131]

Cernich AN, Brennana DM, Barker LM, & Bleiberg J (2007). Sources of error in computerized neuropsychological assessment. Archives of Clinical Neuropsychology, 22(Suppl 1), S39–S48. doi: 10.1016/j.acn.2006.10.004 [PubMed: 17097851]

Cogstate. (2018). Cogstate Pediatric and Adult Normative Data New Haven, CT: Cogstate, Inc.

Cole WR, Arrieux JP, Schwab K, Ivins BJ, Qashu FM, & Lewis SC (2013). Test–retest reliability of four computerized neurocognitive assessment tools in an active duty military population. Archives of Clinical Neuropsychology, 28(7), 732–742. doi: 10.1093/arclin/act040 [PubMed: 23819991]

Collie A, Maruff P, Darby DG, & McStephen M (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. Journal of the International Neuropsychological Society, 9(3), 419–428. doi: 10.1017/S1355617703930074 [PubMed: 12666766]

Cromer JA, Schembri AJ, Harel BT, & Maruff P (2015). The nature and rate of cognitive maturation from late childhood to adulthood. Frontiers in Psychology, 6, 704. doi: 10.3389/fpsyg.2015.00704 [PubMed: 26074853]

Crook TH, Kay GG, & Larrabee GJ (2009). Computer-based cognitive testing. In Grant I & Adams KM (Eds.), Neuropsychological Assessment of Neuropsychiatric and Neuromedical Disorders, Third Edition (pp. 84–100). New York, NY: Oxford University Press, Inc.

Dingwall K, Lewis M, Maruff P, & Cairney S (2010). Assessing cognition following petrol sniffing for Indigenous Australians. The Australian and New Zealand Journal of Psychiatry, 44(7), 631–639. doi: 10.3109/00048671003627405 [PubMed: 20560850]

Fratti S, Bowden SC, & Cook MJ (2017). Reliability and validity of the CogState computerized battery in patients with seizure disorders and healthy young adults: comparison with standard neuropsychological tests. The Clinical Neuropsychologist, 31(3), 569–586. doi: 10.1080/13854046.2016.1256435 [PubMed: 27852143]

Fredrickson J, Maruff P, Woodward M, Moore L, Fredrickson A, Sach J, et al. (2010). Evaluation of the usability of a brief computerized cognitive screening test in older people for epidemiological studies. Neuroepidemiology, 34(2), 65–75. doi: 10.1159/000264823 [PubMed: 20016215]

Gershon R, Nowinski C, Kaat A, & Slotkin J (2017). Equivalence of the NIH Toolbox for assessment of neurological and beahvioral function iPad version. Poster presented at the 45th Annual Meeting of the International Neuropsychological Society, New Orleans, LA.

Hinton-Bayre AD (2010). Deriving reliable change statistics from test-retest normative data: comparison of models and mathematical expressions. Archives of Clinical Neuropsychology, 25(3), 244–256. doi: 10.1093/arclin/acq008 [PubMed: 20197293]

Ivnik RJ, Malec JF, Smith GE, Tangalos EG, Petersen RC, Kokmen E, et al. (1992). Mayo's Older Americans Normative Studies: WAIS-R, WMS-R and AVLT norms for ages 56 through 97. The Clinical Neuropsychologist, 6(Supplement), 1–104.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association; & the National Council on Measurement in Education. (2014). Standards for educational and psychological testing Washington, DC: American Educational Research Association.

Kokmen E, Smith GE, Petersen RC, Tangalos E, & Ivnik RC (1991). The short test of mental status: Correlations with standardized psychometric testing. Archives of Neurology, 48(7), 725–728. [PubMed: 1859300]

Lewis MS, Maruff P, Silbert BS, Evered LA, & Scott DA (2007). The influence of different error estimates in the detection of postoperative cognitive dysfunction using reliable change indices with correction for practice effects. Archives of Clinical Neuropsychology, 22(2), 249–257. doi: 10.1016/j.acn.2006.05.004 , 10.1016/j.acn.2006.05.00410.1016/j.acn.2007.01.002, 10.1016/j.acn.2007.01.002 [PubMed: 17443923]

Lim Y, Ellis K, Harrington K, Ames D, Martins R, Masters C, et al. (2012). Use of the CogState Brief Battery in the assessment of Alzheimer's disease related cognitive impairment in the Australian Imaging, Biomarkers and Lifestyle (AIBL) study. Journal of Clinical and Experimental Neuropsychology, 34(4), 345–358. doi: 10.1080/13803395.2011.643227 [PubMed: 22248010]

Lim YY, Jaeger J, Harrington K, Ashwood T, Ellis KA, Stoffler A, et al. (2013). Three-month stability of the CogState brief battery in healthy older adults, mild cognitive impairment, and Alzheimer's disease: Results from the Australian Imaging, Biomarkers, and Lifestyle-Rate of Change Substudy (AIBL-ROCS). Archives of Clinical Neuropsychology, 28(4), 320–330. doi: 10.1093/arclin/act021 [PubMed: 23552802]

Lim YY, Maruff P, Schindler R, Ott BR, Salloway S, Yoo DC, et al. (2015). Disruption of cholinergic neurotransmission exacerbates Aβ-related cognitive impairment in preclinical Alzheimer's disease. Neurobiology of Aging, 36(10), 2709–2715. doi: 10.1016/j.neurobiolaging.2015.07.009 [PubMed: 26233262]

Louey AG, Cromer JA, Schembri AJ, Darby DG, Maruff P, Makdissi M, et al. (2014). Detecting cognitive impairment after concussion: Sensitivity of change from baseline and normative data methods using the CogSport/Axon cognitive test battery. Archives of Clinical Neuropsychology, 29(5), 432–441. doi: 10.1093/arclin/acu020 [PubMed: 24813184]

MacDonald J, & Duerson D (2015). Reliability of a computerized neurocognitive test in baseline concussion testing of high school athletes. Clinical Journal of Sport Medicine, 25(4), 367–372. doi: 10.1097/JSM.0000000000000139. [PubMed: 25061807]

Maruff P, Lim YY, Darby D, Ellis KA, Pietrzak RH, Snyder PJ, et al. (2013). Clinical utility of the cogstate brief battery in identifying cognitive impairment in mild cognitive impairment and Alzheimer's disease. BMC Psychology, 1(1), 30. doi: 10.1186/2050-7283-1-30 [PubMed: 25566378]

Mielke MM, Machulda MM, Hagen CE, Christianson TJ, Roberts RO, Knopman DS, et al. (2016). Influence of amyloid and APOE on cognitive performance in a late middle-aged cohort. Alzheimer's & Dementia, 12(3), 281–291. doi: 10.1016/j.jalz.2015.09.010

Mielke MM, Machulda MM, Hagen CE, Edwards KK, Roberts RO, Pankratz VS, et al. (2015). Performance of the CogState computerized battery in the Mayo Clinic Study on Aging. Alzheimer's & Dementia, 11(11), 1367–1376. doi: 10.1016/j.jalz.2015.01.008

Mielke MM, Weigand SD, Wiste HJ, Vemuri P, Machulda MM, Knopman DS, et al. (2014). Independent comparison of CogState computerized testing and a standard cognitive battery with neuroimaging. Alzheimer's & Dementia, 10(6), 779–789. doi: 10.1016/j.jalz.2014.09.001

Morris JC (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology, 43(11), 2412–2414.

Nelson LD, LaRoche AA, Pfaller AY, Lerner EB, Hammeke TA, Randolph C, et al. (2016). Prospective, head-to-head study of three computerized neurocognitive assessment tools (CNTs): Reliability and validity for the assessment of sport-related concussion. Journal of the International Neuropsychological Society, 22(1), 24–37. doi: 10.1017/S1355617715001101 [PubMed: 26714883]

Parsons TD, McMahan T, & Kane R (2018). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. The Clinical Neuropsychologist, 32(1), 16–41. doi: 10.1080/13854046.2017.1337932 [PubMed: 28590154]

Petersen RC, Roberts RO, Knopman DS, Geda YE, Cha RH, Pankratz VS, et al. (2010). Prevalence of mild cognitive impairment is higher in men: The Mayo Clinic Study of Aging. Neurology, 75(10), 889–897. doi: 10.1212/WNL.0b013e3181f11d85 [PubMed: 20820000]

Pietrzak RH, Maruff P, Mayes LC, Roman SA, Sosa JA, & Snyder PJ (2008). An examination of the construct validity and factor structure of the Groton Maze Learning Test, a new measure of spatial working memory, learning efficiency, and error monitoring. Archives of Clinical Neuropsychology, 23(4), 433–445. doi: 10.1016/j.acn.2008.03.002 [PubMed: 18448309]

Rabin LA, Spadaccini AT, Brodale DL, Grant KS, Elbulok-Charcape MM, & Barr WB (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. Professional Psychology: Research and Practice, 45(5), 368–377. doi: 10.1037/a0037987

Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, et al. (2008). The Mayo Clinic Study of Aging: Design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology, 30(1), 58–69. doi: 10.1159/000115751 [PubMed: 18259084]

Salmon JP, Jones SA, Wright CP, Butler BC, Klein RM, & Eskes GA (2017). Methods for validating chronometry of computerized tests. Journal of Clinical and Experimental Neuropsychology, 39(2), 190–210. doi: 10.1080/13803395.2016.1215411 [PubMed: 27532256]

Slick DJ (2006). Psychometrics in neuropsychological assessment. In Strauss E, Sherman EMS & Spreen O (Eds.), A Compendium of Neuropsychological Tests: Administration, Norms and Commentary New York, NY: Oxford University Press.

Snyder PJ, Bednar MM, Cromer JR, & Maruff P (2005). Reversal of scopolamine-induced deficits with a single dose of donepezil, an acetylcholinesterase inhibitor. Alzheimer's & Dementia, 1(2), 126–135. doi: 10.1016/j.jalz.2005.09.004

Snyder PJ, Jackson CE, Petersen RC, Khachaturian AS, Kaye J, Albert MS, et al. (2011). Assessment of cognition in mild cognitive impairment: A comparative study. Alzheimer's & Dementia, 7(3), 338–355. doi: 10.1016/j.jalz.2011.03.009

Valdes EG, Sadeq NA, Harrison Bush A. L., Morgan D, & Andel R (2016). Regular cognitive self-monitoring in community-dwelling older adults using an internet-based tool. Journal of Clinical and Experimental Neuropsychology, 38(9), 1026–1037. doi: 10.1080/13803395.2016.1186155 [PubMed: 27266359]

Wild K, Howieson D, Webbe F, Seelye A, & Kaye J (2008). Status of computerized cognitive testing in aging: a systematic review. Alzheimer's & Dementia, 4(6), 428–437. doi: 10.1016/j.jalz.2008.07.003

Zygouris S, & Tsolaki M (2015). Computerized cognitive testing for older adults: A review. American Journal of Alzheimer's Disease and Other Dementias, 30(1), 13–28. doi: 10.1177/1533317514522852
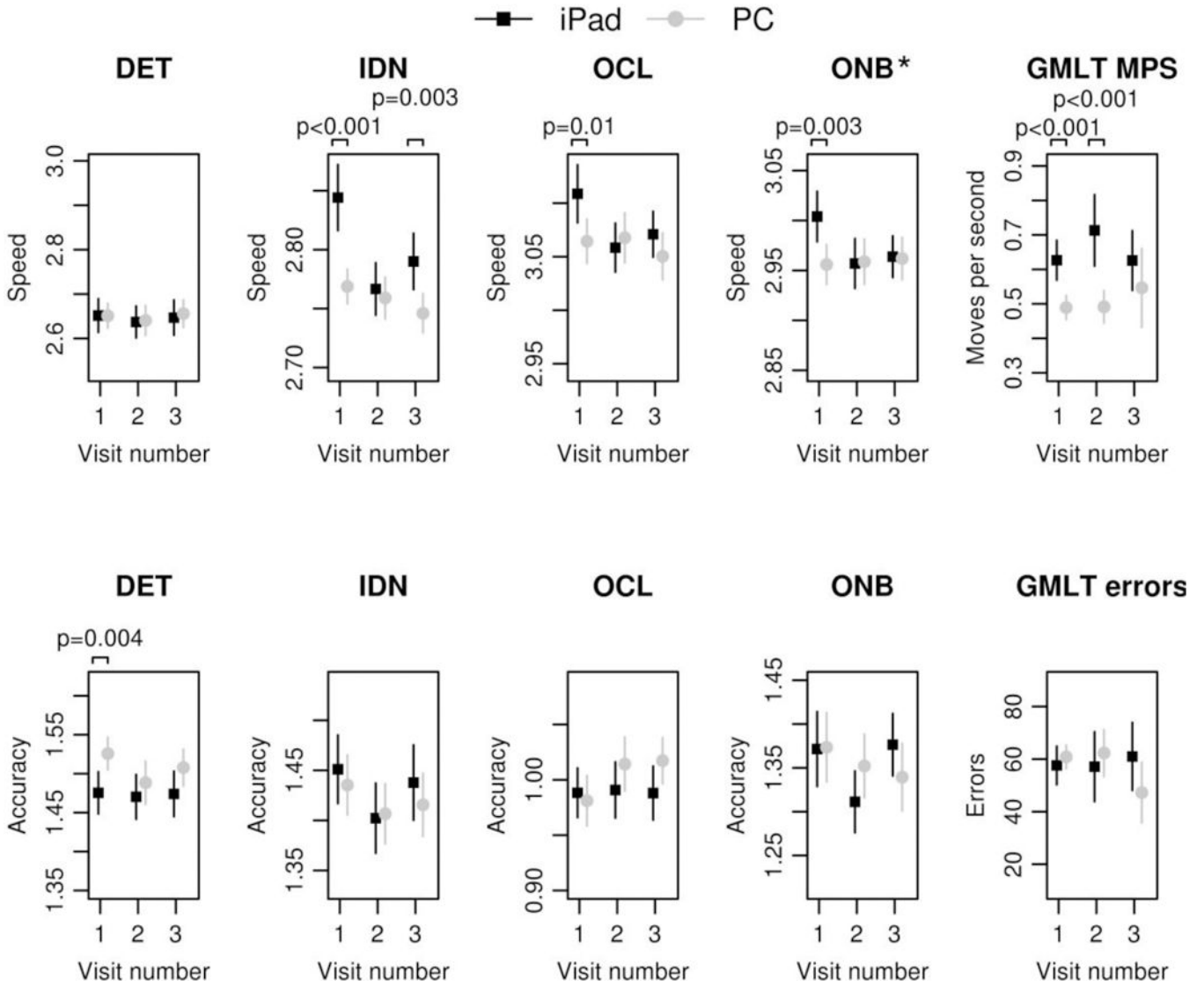
Figure 1.
Means (and standard deviation bars) for each time point for longitudinal sample by PC and iPad.

Note. PC = personal computer; SD = standard deviation; DET = Detection; IDN = Identification; OCL = One Card Learning; ONB = One Back; GMLT = Groton Maze Learning Test; MPS = Moves Per Second. * *p* < .05.

Table 1.

Test-retest sample demographics and frequency of completion and integrity failures.

	iPad (N=168)	PC (N=168)	Total	*p*
Demographics			(N=168)	
Age Mean (*SD*)			64.4 (5.39)	
Years of Education Mean (*SD*)			15.2 (2.27)	
Percent Male			52%	
Prior Cogstate Visits Mean (*SD*)			2.35 (1.09)	
Completion, n (%) failed			(N=336)	
Detection	0 (0.0%)	1 (0.6%)	1 (0.3%)	0.32
Identification	0 (0.0%)	1 (0.6%)	1 (0.3%)	0.32
One Card Learning	1 (0.6%)	2 (1.2%)	3 (0.9%)	0.56
One Back	0 (0.0%)	2 (1.2%)	2 (0.6%)	0.16
Groton Maze Learning Test	0 (0.0%)	1 (0.6%)	1 (0.3%)	0.32
Any completion flag failed	1 (0.6%)	4 (2.4%)	5 (1.5%)	0.18
Integrity, n (%) failed			(N=336)	
Detection	6 (3.6%)	2 (1.2%)	8 (2.4%)	0.15
Identification	1 (0.6%)	4 (2.4%)	5 (1.5%)	0.18
One Card Learning	2 (1.2%)	1 (0.6%)	3 (0.9%)	0.56
One Back	2 (1.2%)	1 (0.6%)	3 (0.9%)	0.56
Groton Maze Learning Test	0 (0.0%)	2 (1.2%)	2 (0.6%)	0.16
Any integrity flag failed	11 (6.5%)	9 (5.4%)	20 (6.0%)	0.64

Note. Overall failures were counted once even if more than one portion of Cogstate was failed for that participant's visit. Thus overall count may differ from the sum of the parts. PC = personal computer; SD = standard deviation. *P*-values reported above are from linear model ANOVAs (continuous variables) or Pearson's Chi-square test (frequencies).

Table 2.

Test naïve sample demographics and frequency of completion and integrity failures.

	iPad	PC	*p*
Demographics – Brief Battery	(n = 605)	(n = 1820)	
Age Mean (*SD*)	76.3 (11.8)	66.3 (9.58)	<.001
Years of Education Mean (*SD*)	14.6 (2.61)	15.1 (2.4)	<.001
Percent Male	50%	50%	0.699
Demographics – GMLT	(n = 195)	(n = 1634)	
Age Mean (*SD*)	62.8 (9.77)	64.9 (8.79)	0.002
Education Mean (*SD*)	15.1 (2.28)	15.1 (2.37)	0.895
Percent Male	49%	51%	0.428
Completion, n (%) failed			
Detection	16 (2.6%)	0 (0.0%)	<.001
Identification	16 (2.7%)	1 (0.1%)	<.001
One Card Learning	8 (1.3%)	3 (0.2%)	<.001
One Back	16 (2.7%)	1 (0.1%)	<.001
Groton Maze Learning Test	5 (2.5%)	25 (1.5%)	0.292
Any completion flag failed	29 (4.8%)	27 (1.5%)	<.001
Integrity, n (%) failed			
Detection	54 (9.0%)	56 (3.1%)	<.001
Identification	32 (5.3%)	28 (1.5%)	<.001
One Card Learning	56 (9.3%)	37 (2.0%)	<.001
One Back	59 (9.9%)	49 (2.7%)	<.001
Groton Maze Learning Test	7 (3.5%)	13 (0.8%)	0.004
Any integrity flag failed	130 (21.5%)	149 (8.2%)	<.001

Note. Overall failures were counted once even if more than one portion of Cogstate was failed for that participant's visit. Thus overall count may differ from the sum of the parts. PC = personal computer; SD = standard deviation. *P*-values reported above are from linear model ANOVAs (continuous variables) or Pearson's Chi-square test (frequencies).

Table 3.

Baseline demographics and frequency of completion and integrity failures across all visits for the longitudinal sample.

	iPad	PC	*p*
Demographics – Brief Battery	(n = 63)	(n = 63)	
Age Mean (*SD*)	79.4 (5.43)	78.8 (5.46)	0.58
Years of Education Mean (*SD*)	14.3 (2.22)	14.5 (2.49)	0.74
Percent Male	57%	57%	1.00
Duration of follow-up (months)	16.2 (2.68)	17.3 (2.74)	0.02
Demographics – GMLT	(n = 25)	(n = 62)	
Age Mean (*SD*)	75.6 (3.47)	78.6 (5.24)	0.009
Education Mean (*SD*)	15.4 (2.04)	14.5 (2.49)	0.11
Percent Male	60%	57%	0.95
Completion, n (%) failed			
Detection	2 (1%)	0 (0%)	0.48
Identification	2 (1%)	0 (0%)	0.48
One Card Learning	0 (0%)	1 (1%)	1.00
One Back	1 (1%)	0 (0%)	1.00
Groton Maze Learning Test	0 (0%)	5 (5%)	0.29
Any integrity flag failed	2 (1%)	5 (3%)	0.45
Integrity, n (%) failed			
Detection	12 (6%)	14 (7%)	0.84
Identification	10 (5%)	3 (2%)	0.09
One Card Learning	9 (5%)	5 (3%)	0.42
One Back	4 (2%)	4 (2%)	1.00
Groton Maze Learning Test	1 (2%)	1 (1%)	1.00
Any integrity flag failed	31 (16%)	25 (13%)	0.47

Note. Overall failures were counted once even if more than one portion of Cogstate was failed for that participant's visit. Thus overall count may differ from the sum of the parts. PC = personal computer; SD = standard deviation. *P*-values reported above are from linear model ANOVAs (continuous variables) or Pearson's Chi-square test (frequencies).

Table 4.

Comparison of iPad and PC Cogstate results administered the same day (test-retest sample).

	N	Clinic iPad mean (SD)	Clinic PC mean (SD)	Mean Difference iPad – PC mean (SD)	Mean Difference 95% CI	t	p-value	d
DET speed	159	2.612 (0.119)	2.591 (0.108)	0.024 (0.102)	0.008, 0.040	3.00	0.003	−0.24 (−0.46, −0.02)
DET raw speed		427.075 (144.638)	403.336 (113.153)	26.607 (132.891)	5.792, 47.423	2.52	0.013	−0.20 (−0.42, 0.02)
DET variability		0.087 (0.036)	0.088 (0.038)	−0.002 (0.042)	−0.008 ,0.005	−0.56	0.580	0.04 (−0.18, 0.27)
DET accuracy		1.513 (0.094)	1.514 (0.09)	−0.005 (0.114)	−0.023 ,0.013	−0.54	0.593	0.04 (−0.18, 0.26)
DET raw accuracy		0.988 (0.022)	0.989 (0.02)	−0.002 (0.026)	−0.006 ,0.002	−0.86	0.392	0.07 (−0.15, 0.29)
IDN speed	163	2.784 (0.082)	2.723 (0.064)	0.025 (0.064)	0.015, 0.035	4.96	<.001	−0.39 (−0.61, −0.17)
IDN raw speed		569.995 (117.63)	534.51 (81.104)	35.958 (91.175)	21.856, 50.06	5.04	<.001	−0.39 (−0.61, −0.17)
IDN variability		0.082 (0.022)	0.081 (0.022)	0.001 (0.027)	−0.003 ,0.006	0.68	0.498	−0.05 (−0.27, 0.17)
IDN accuracy		1.461 (0.117)	1.47 (0.113)	−0.007 (0.153)	−0.031 ,0.017	−0.59	0.558	0.07 (−0.17, 0.26)
IDN raw accuracy		0.975 (0.03)	0.977 (0.029)	−0.002 (0.040)	−0.008 ,0.004	−0.70	0.486	0.06 (−0.16, 0.27)
OCL speed	163	3.021 (0.076)	3.010 (0.710)	0.01 (0.051)	0.002, 0.017	2.42	0.016	−0.19 (−0.41, 0.03)
OCL raw speed		1066.006 (199.155)	1037.028 (182.157)	25.11 (129.100)	5.142, 45.078	2.48	0.014	−0.19 (−0.41, 0.02)
OCL variability		0.133 (0.034)	0.136 (0.033)	−0.004 (0.034)	−0.010 ,0.001	−1.61	0.109	0.13 (−0.09, 0.34)
OCL accuracy		1.029 (0.086)	1.039 (0.084)	−0.009 (0.103)	−0.025, 0.007	−1.13	0.261	0.09 (−0.13, 0.31)
OCL raw accuracy		0.731 (0.075)	0.74 (0.073)	−0.008 (0.090)	−0.022 ,0.006	−1.14	0.255	0.09 (−0.13, 0.31)
ONB speed	163	2.915 (0.087)	2.890 (0.078)	0.026 (0.063)	0.016, 0.035	5.17	<.001	−0.41 (−0.63, −0.18)
ONB raw speed		839.169 (172.751)	788.903 (144.592)	51.314 (128.131)	31.495, 71.132	5.11	<.001	−0.40 (−0.62, −0.18)
ONB variability		0.12 (0.033)	0.125 (0.028)	−0.005 (0.032)	−0.010, 0.000	−1.96	0.052	0.15 (−0.07, 0.37)
ONB accuracy		1.419 (0.126)	1.417 (0.125)	0.001 (0.147)	−0.021, 0.024	0.12	0.901	−0.01 (−0.23, 0.21)
ONB raw accuracy		0.962 (0.04)	0.962 (0.044)	−0.001 (0.046)	−0.008 ,0.006	−0.21	0.834	0.02 (−0.20, 0.23)
GMLT MPS	160	0.737 (0.158)	0.632 (0.135)	0.109 (0.132)	0.088, 0.129	10.42	<.001	−0.82 (−1.05, −0.59)
GMLT total errors		48.327 (16.348)	44.344 (13.983)	3.244 (14.599)	0.964, 5.523	2.81	0.006	−0.22 (−0.44, 0.00)

Note. Higher values for speed variables indicate slower response time; higher GMLT MPS indicate faster response time. PC = personal computer; SD = standard deviation; DET = Detection; IDN = Identification; OCL = One Card Learning; ONB = One Back; GMLT = Groton Maze Learning Test; MPS = Moves Per Second.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5.

Reliability coefficients: Pearson's rho (ICC) for test-retest sample and longitudinal sample (baseline to 7.5 months, 7.5 to 15 months).

	Test-retest interval					
	Same Day	**Baseline - 7.5 months**		**7.5 – 15 months**		
	PC-iPad	**PC-PC**	**iPad-iPad**	**PC-PC**	**iPad-iPad**	
DET speed	0.59 (0.57)	0.63 (0.61)	0.45 (0.44)	0.64 (0.64)	0.50 (0.50)	
IDN speed	0.64 (0.58)	0.55 (0.53)	0.49 (0.47)	0.57 (0.56)	0.52 (0.52)	
OCL speed	0.77 (0.76)	0.72 (0.72)	0.64 (0.63)	0.74 (0.74)	0.73 (0.73)	
OCL accuracy	0.26 (0.26)	0.36 (0.36)	0.51 (0.51)	0.50 (0.49)	0.59 (0.59)	
ONB speed	0.71 (0.67)	0.76 (0.75)	0.60 (0.60)	0.73 (0.73)	0.72 (0.71)	
ONB accuracy	0.30 (0.30)	0.32 (0.32)	0.16 (0.16)	0.43 (0.43)	0.22 (0.22)	
GMLT MPS	0.59 (0.40)	0.80 (0.80)	0.82 (0.82)	0.79 (0.79)	0.52 (0.52)	
GMLT total errors	0.52 (0.50)	0.84 (0.84)	0.72 (0.72)	0.28 (0.25)	0.46 (0.46)	

Note. PC = personal computer; SD = standard deviation; DET = Detection; IDN = Identification; OCL = One Card Learning; ONB = One Back; GMLT = Groton Maze Learning Test; MPS = Moves Per Second. Sample sizes for each measure and each time point are provided in Tables 4 and 7.

Table 6.

Linear regression model estimates (standard errors) for the test naïve sample.

	N	iPad vs. PC Group Difference Estimate[1] *(std. error)*	Age	Non-Linear Age	Education	Sex
			Model Covariate Estimate[1] *(std. error)*			
DET speed	2303	0.004 (0.005)	0.003**(0.001)	0.001 (0.001)	−0.005**(0.001)	−0.017**(0.004)
DET raw speed		14.044*(6.173)	2.832**(0.598)	1.382 (0.813)	−5.171**(0.965)	−16.896**(4.721)
DET accuracy		−0.024**(0.005)	0.000 (0.000)	0.000 (0.001)	0.005**(0.001)	0.000 (0.004)
DET raw accuracy		−0.005**(0.001)	0.000 (0.000)	0.000 (0.000)	0.001**(0.000)	0.000 (0.001)
IDN speed	2352	0.057**(0.004)	0.002**(0.000)	0.000 (0.001)	−0.003**(0.001)	0.001 (0.003)
IDN raw speed		91.180**(5.822)	2.608**(0.568)	0.413 (0.771)	−4.601**(0.920)	1.459 (4.473)
IDN accuracy		0.032**(0.007)	−0.001 (0.001)	−0.001 (0.001)	0.002 (0.001)	−0.015*(0.005)
IDN raw accuracy		0.007**(0.002)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	−0.005*(0.002)
OCL speed	2321	0.043**(0.005)	0.002**(0.000)	0.001 (0.001)	0.001 (0.001)	0.000 (0.004)
OCL raw speed		132.010**(15.109)	4.241*(1.471)	5.257*(1.993)	1.040 (2.386)	−2.581 (11.603)
OCL accuracy	2321	0.001(0.005)	−0.003**(0.000)	0.002*(0.001)	0.004**(0.001)	0.003 (0.004)
OCL raw accuracy		0.001 (0.004)	−0.002**(0.000)	0.001*(0.001)	0.003**(0.001)	0.003 (0.003)
ONB speed	2297	0.029**(0.005)	0.003**(0.000)	0.000 (0.001)	−0.003**(0.001)	−0.003 (0.004)
ONB raw speed		66.622**(10.565)	4.987**(1.023)	1.984 (1.405)	−5.898**(1.665)	−8.158 (8.079)
ONB accuracy	2297	0.027**(0.008)	−0.001 (0.001)	−0.002 (0.001)	0.005**(0.001)	−0.020*(0.006)
ONB raw accuracy		0.007*(0.003)	0.000 (0.000)	−0.001*(0.000)	0.002**(0.001)	−0.007*(0.003)
GMLT MPS	1814	0.117**(0.010)	−0.009**(0.001)	0.002 (0.001)	0.007**(0.001)	0.020*(0.006)
GMLT total errors		3.514*(1.297)	0.439**(0.096)	0.258 (0.148)	−1.025**(0.168)	−4.977**(0.796)

Note. Higher values for speed variables indicate slower response time; higher GMLT MPS indicate faster response time. PC = personal computer; SD = standard deviation; DET = Detection; IDN = Identification; OCL = One Card Learning; ONB = One Back; GMLT = Groton Maze Learning Test; MPS = Moves Per Second.

[1]
* $p < .05$;** $p < .001$

Table 7.

Comparison of iPad and PC groups at each time point (longitudinal sample).

	N iPad	Clinic iPad mean (SD)	N PC	Clinic PC Mean (SD)	Mean Difference iPad – PC mean (95% CI)	p-value	d
BASELINE							
DET speed	57	2.65 (0.14)	62	2.65 (0.11)	0.00 (−0.046, 0.047)	0.98	0.00
DET raw speed		477.42 (209)		463.14 (136)	14.27 (−50, 79)	0.66	−0.08
DET variability		0.11 (0.045)		0.08 (0.032)	0.03 (0.019, 0.047)	<0.001	−0.85
DET accuracy		1.48 (0.1)		1.53 (0.082)	−0.05 (−0.084, −0.017)	0.004	0.54
DET raw accuracy		0.98 (0.023)		0.99 (0.017)	−0.01 (−0.018, −0.0031)	0.006	0.52
IDN speed	60	2.84 (0.11)	63	2.77 (0.058)	0.08 (0.044, 0.11)	<0.001	−0.86
IDN raw speed		719.86 (175)		592.54 (83)	127.32 (78, 177)	<0.001	−0.92
IDN variability		0.09 (0.026)		0.09 (0.022)	0.00 (−0.0066, 0.011)	0.65	−0.08
IDN accuracy		1.45 (0.13)		1.44 (0.12)	0.02 (−0.029, 0.061)	0.49	−0.12
IDN raw accuracy		0.97 (0.04)		0.97 (0.033)	0.00 (−0.012, 0.014)	0.90	−0.02
OCL speed	59	3.11 (0.1)	61	3.06 (0.081)	0.04 (0.011, 0.078)	0.01	−0.48
OCL raw speed		1320.54 (308)		1179.69 (223)	140.85 (43, 238)	0.005	−0.52
OCL variability		0.15 (0.038)		0.14 (0.028)	0.01 (−0.0048, 0.019)	0.24	−0.22
OCL accuracy		0.99 (0.086)		0.98 (0.089)	0.01 (−0.024, 0.039)	0.65	−0.08
OCL raw accuracy		0.69 (0.077)		0.69 (0.08)	0.01 (−0.022, 0.035)	0.64	−0.09
ONB speed	60	3.00 (0.098)	61	2.96 (0.077)	0.05 (0.016, 0.08)	0.003	−0.55
ONB raw speed		1034.17 (227)		917.82 (167)	116.35 (44, 188)	0.002	−0.58
ONB variability		0.13 (0.03)		0.14 (0.033)	0.00 (−0.013, 0.0093)	0.72	0.06
ONB accuracy		1.37 (0.17)		1.37 (0.15)	0.00 (−0.06, 0.056)	0.95	0.01
ONB raw accuracy		0.94 (0.065)		0.94 (0.064)	0.00 (−0.027, 0.019)	0.73	0.06
GMLT MPS	24	0.63 (0.14)	59	0.49 (0.13)	0.14 (0.071, 0.2)	<0.001	−1.01
GMLT Total Errors		57.58 (17)		60.83 (17)	−3.25 (−12, 5.2)	0.44	0.19
TIME 2							
DET speed	57	2.64 (0.13)	54	2.64 (0.12)	0.00 (−0.052, 0.045)	0.89	0.03
DET raw speed		456.85 (170)		456.82 (156)	0.03 (−61, 61)	>0.99	0.00
DET variability		0.09 (0.026)		0.10 (0.045)	−0.01 (−0.027, 0.00058)	0.06	0.36

	N iPad	Clinic iPad mean (SD)	N PC	Clinic PC Mean (SD)	Mean Difference iPad – PC mean (95% CI)	p-value	d
DET accuracy		1.47 (0.11)		1.49 (0.1)	-0.02 (-0.058, 0.021)	0.36	0.17
DET raw accuracy		0.98 (0.026)		0.98 (0.023)	0.00 (-0.014, 0.0045)	0.31	0.19
IDN speed	59	2.77 (0.085)	62	2.76 (0.07)	0.01 (-0.02, 0.036)	0.59	-0.10
IDN raw speed		595.71 (121)		581.89 (104)	13.81 (-27, 55)	0.50	-0.12
IDN variability		0.09 (0.022)		0.09 (0.027)	0.00 (-0.0067, 0.011)	0.64	-0.08
IDN accuracy		1.40 (0.13)		1.41 (0.12)	0.00 (-0.05, 0.041)	0.85	0.03
IDN raw accuracy		0.96 (0.047)		0.96 (0.039)	-0.01 (-0.021, 0.01)	0.51	0.12
OCL speed	59	3.06 (0.088)	61	3.07 (0.091)	-0.01 (-0.042, 0.023)	0.57	0.10
OCL raw speed		1167.97 (250)		1195.09 (266)	-27.13 (-120, 66)	0.57	0.11
OCL variability		0.14 (0.035)		0.15 (0.047)	-0.01 (-0.025, 0.0054)	0.20	0.23
OCL accuracy		0.99 (0.097)		1.01 (0.095)	-0.02 (-0.058, 0.011)	0.18	0.24
OCL raw accuracy		0.70 (0.086)		0.72 (0.083)	-0.02 (-0.052, 0.0093)	0.17	0.25
ONB speed	61	2.96 (0.098)	62	2.96 (0.089)	0.00 (-0.035, 0.032)	0.91	0.02
ONB raw speed		928.86 (216)		929.67 (212)	-0.82 (-77, 76)	0.98	0.00
ONB variability		0.13 (0.029)		0.14 (0.036)	-0.01 (-0.023, 0.00043)	0.06	0.34
ONB accuracy		1.31 (0.14)		1.35 (0.14)	-0.04 (-0.091, 0.0091)	0.11	0.29
ONB raw accuracy		0.92 (0.066)		0.94 (0.061)	-0.02 (-0.04, 0.0057)	0.14	0.27
GMLT MPS	10	0.71 (0.14)	27	0.49 (0.12)	0.22 (0.11, 0.33)	<0.001	-1.60
GMLT Total Errors		57.10 (19)		62.30 (23)	-5.20 (-20, 10)	0.49	0.26
TIME 3							
DET speed	61	2.65 (0.15)	59	2.66 (0.12)	-0.01 (-0.059, 0.041)	0.73	0.06
DET raw speed		478.20 (235)		471.02 (146)	7.18 (-63, 78)	0.84	-0.04
DET variability		0.10 (0.039)		0.10 (0.041)	0.00 (-0.017, 0.012)	0.77	0.05
DET accuracy		1.47 (0.11)		1.51 (0.09)	-0.03 (-0.071, 0.0033)	0.07	0.33
DET raw accuracy		0.98 (0.028)		0.99 (0.018)	-0.01 (-0.018, -0.0013)	0.02	0.42
IDN speed	57	2.79 (0.09)	61	2.75 (0.065)	0.04 (0.015, 0.073)	0.003	-0.56
IDN raw speed		629.96 (134)		563.54 (89)	66.41 (24, 108)	0.002	-0.58
IDN variability		0.09 (0.028)		0.09 (0.022)	0.00 (-0.0075, 0.011)	0.72	-0.07
IDN accuracy		1.44 (0.14)		1.42 (0.12)	0.02 (-0.026, 0.071)	0.36	-0.17
IDN raw accuracy		0.96 (0.049)		0.96 (0.036)	0.00 (-0.014, 0.018)	0.82	-0.04

	N iPad	Clinic iPad mean (SD)	N PC	Clinic PC Mean (SD)	Mean Difference iPad − PC mean (95% CI)	p-value	d
OCL speed	63	3.07 (0.084)	61	3.05 (0.085)	0.02 (−0.0096, 0.051)	0.18	−0.24
OCL raw speed		1199.98 (245)		1144.18 (224)	55.80 (−28, 140)	0.19	−0.24
OCL variability		0.14 (0.037)		0.15 (0.035)	0.00 (−0.015, 0.01)	0.70	0.07
OCL accuracy		0.99 (0.096)		1.02 (0.081)	−0.03 (−0.061, 0.0025)	0.07	0.33
OCL raw accuracy		0.69 (0.087)		0.72 (0.071)	−0.03 (−0.055, 0.0011)	0.06	0.34
ONB speed	62	2.96 (0.083)	62	2.96 (0.083)	0.00 (−0.027, 0.031)	0.90	−0.02
ONB raw speed		936.05 (172)		932.91 (183)	3.14 (−60, 66)	0.92	−0.02
ONB variability		0.13 (0.03)		0.14 (0.031)	−0.01 (−0.017, 0.0044)	0.24	0.21
ONB accuracy		1.38 (0.14)		1.34 (0.15)	0.04 (−0.015, 0.089)	0.16	−0.25
ONB raw accuracy		0.94 (0.055)		0.93 (0.071)	0.02 (−0.0054, 0.039)	0.13	−0.27
GMLT MPS	11	0.63 (0.13)	8	0.55 (0.14)	0.08 (−0.053, 0.21)	0.22	0.60
GMLT Total Errors		61.00 (19)		47.25 (19)	13.75 (−2.2, 30)	0.09	0.80

Note. Higher values for speed variables indicate slower response time; higher GMLT MPS indicate faster response time. PC = personal computer; SD = standard deviation; DET = Detection; IDN = Identification; OCL = One Card Learning; ONB = One Back; GMLT = Groton Maze Learning Test; MPS = Moves Per Seconds.