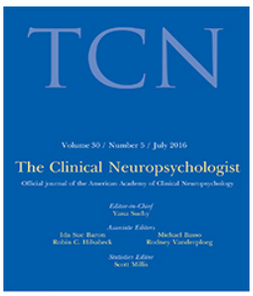

The Clinical Neuropsychologist

ISSN: 1385-4046 (Print) 1744-4144 (Online) Journal homepage: https://www.tandfonline.com/loi/ntcn20

Digital neuropsychology: challenges and opportunities at the intersection of science and software

Laura Germine, Katharina Reinecke & Naomi S. Chaytor

To cite this article: Laura Germine, Katharina Reinecke & Naomi S. Chaytor (2019): Digital neuropsychology: challenges and opportunities at the intersection of science and software, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2018.1535662

To link to this article: https://doi.org/10.1080/13854046.2018.1535662

THE CLINICAL NEUROPSYCHOLOGIST
https://doi.org/10.1080/13854046.2018.1535662

Digital neuropsychology: challenges and opportunities at the intersection of science and software

Laura Germine[a,b,c], Katharina Reinecke[d] and Naomi S. Chaytor[e]

[a]Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA; [b]Department of Psychiatry, Harvard Medical School, Boston, MA, USA; [c]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA; [d]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA; [e]Elson S. Floyd College of Medicine, Washington State University, Spokane, WA, USA

ABSTRACT

Objective: Digital devices are now broadly accessible and have the capacity to measure aspects of human behavior with high precision and accuracy, in a standardized manner. The purpose of this article is to characterize opportunities and barriers for modern digital neuropsychology, particularly those that are unique to digital assessment.

Methods: We provide a critical overview of the state-of-the-art in digital neuropsychology, focusing on personal digital devices.

Results: We identify three major barriers associated with digital neuropsychology, which affect both the interpretation of test scores and test norms: (1) variability in the perceptual, motor and cognitive demands of the same test across digital device classes (e.g. personal computer, tablet and smartphone); (2) hardware and software variability between devices within the same class that affect stimulus presentation and measurement and (3) rapid changes over time in hardware, software and device ownership, which can lead to rapid obsolescence of particular tests and test norms. We offer specific recommendations to address these barriers and outline new opportunities to understand and measure neuropsychological functioning over time and in everyday environments.

Conclusions: Digital neuropsychology provides new approaches for measuring and monitoring neuropsychological functioning, informed by an understanding of the limitations and potential of digital technology.

ARTICLE HISTORY
Received 17 April 2018
Accepted 27 September 2018

KEYWORDS
Mobile health; web-based assessment; digital neuropsychology; computerized assessment; digital technology

Introduction

Digital approaches to neuropsychological assessment have many recognized advantages in terms of accessibility, richness of measurement, standardization and cost. With the widespread adoption of digital technologies across clinics, research laboratories, and by patients and participants themselves, major shifts towards use of such

CONTACT Laura Germine lgermine@mclean.harvard.edu Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA

tools to assess cognition and behavior are already occurring (e.g. NIH Toolbox and Cogstate C3: Buckley et al., 2017; Wechsler Q-Interactive: Daniel, Wahlstrom, & Zhang, 2014). Digital technology also offers a level of precise stimulus control and behavior measurement that is difficult or impossible to achieve with traditional paper-and-pencil-based neuropsychological assessment. Such precision allows measures to be adapted from experimental psychology and cognitive neuroscience, including those that require parametric stimulus control and precise quantification of reactions times and changes in reaction times that occur over seconds or minutes (De Leeuw, 2015; Germine et al., 2012; Reimers & Stewart, 2015). Personal digital devices and wearables are also equipped with increasingly sophisticated audiovisual and sensor technology (Onnela & Rauch, 2016), permitting the collection of primary and secondary data about behavior that can supplement traditional neuropsychological assessment (Giannouli, Bock, & Zijlstra, 2018; Min et al., 2014) and, in some cases, may provide data that obviate the need for more burdensome assessments and provide greater ecological validity (Insel, 2017). With decreasing costs and broader adoption of digital devices in the population, worldwide, the idea that data from such devices will be integrated into clinical assessment is almost inevitable. The question, therefore, is how neuropsychologists, psychometricians, research scientists and clinicians should guide such changes towards developing a robust science and practice of digital neuropsychology for research and clinical care.

What is digital neuropsychology?

We introduce the term *digital neuropsychology* to refer to the neuropsychological assessment of cognition and behavior using digital tools, which includes desktop and laptop computers, as well as modern smartphones, tablet computers and wearable devices. Digital neuropsychology is *not* simply the substitution of paper and pencil for a computer screen and electronic response capture; rather, it is a shift in the way we conceptualize neuropsychological measurement that encapsulates both the challenges of digital assessment as well as the opportunities. This means, simultaneously, a shift towards developing and incorporating more sophisticated models of behavior that emphasize the sorts of moment-to-moment data that can be easily captured with digital devices (e.g. variability in reaction time within a test; Hultsch & MacDonald, 2004) as well as accounting for the potential confounds that come with digital assessment (e.g. differences in input latency, described below; Koudritzky et al., 2017).

Challenges in digital neuropsychology

Here we focus on issues related to digital devices themselves and the relationship between users and devices. We omit potential challenges related to digital neuropsychological assessment such as the role of a trained administrator, the influence of testing context (e.g. environmental distractions during in-home smartphone-based assessment) and issues related to data privacy/security. While these other issues are consequential, they are outside the scope of this report and discussed elsewhere (see Kane & Parsons, 2017). We further focus our analysis on the use of personal digital

devices in neuropsychological assessment—that is, digital devices such as tablets and smartphones where we have observed the most dramatic shifts in adoption across demographics, as these devices are both the next frontier of digital neuropsychology and the most poorly understood due to their relative newness as assessment tools. We note that such a focus omits emerging technologies, such as vitual reality (VR), that are currently being adapted for digital neuropsychology with promising applications (e.g. Díaz-Orueta et al., 2014; Iriarte et al., 2016), but are not yet widely accessible across demographic groups.

Challenge 1: changes in test format can impact function

The translation of a paper-and-pencil test to digital format typically necessitates changes to the perceptual, cognitive and/or motor complexity of a task. In some cases, the basic construct being measured by a task remains the same. In other cases, these changes in complexity or response modality may substantially threaten task validity (Bailey, Neigel, Dhanani, & Sims, 2017; Woodward et al., 2016). Consider the basic motor requirements of a trail making test (Crowe, 1998; Reitan, 1958), which requires the user to connect a set of circles with numbers (Part A) or numbers and letters (alternating; Part B) in ascending order (Corrigan & Hinkeldey, 1987). The dependent measure is how long it takes the participant to connect all the circles, in the appropriate order. Successful completion of a trail making test requires the patient/participant to hold and manipulate a pencil in the traditional format, map movements of a mouse on a tabletop to movements of a cursor on a computer screen in the case of computerized assessment with a traditional external mouse, and the movement of a fingertip across the screen with a touchscreen. For touchscreens, such movements are shorter for smaller screens and longer for larger screens, with greater hand occlusion of the stimulus on smaller screens than larger screens. These very different perceptual and motor challenges might translate to better or worse performance for each part of this task.

Based on the analysis of data from 8304 participants aged 18–35 years tested on the citizen science research website TestMyBrain.org (40% female), the same part of a digital trail making test takes 24–30% less time to complete on an iPad than on a Macintosh personal computer and 30–31% more time on an iPhone. Comparing within device classes (e.g. iOS vs. Android), there is a less than 1% difference in completion time. Perhaps more critical, the correlation in performance between the two major parts of the test also differs by device type. Shared variance between performance on Trail A (numbers only) vs. Trail B (numbers and letters) is 140% higher when the test was completed on a tablet vs. a smartphone, representing the impact of screen size, and 50–90% higher on a tablet vs. desktop/laptop, representing the impact of input type. Again, differences within device classes in terms of shared variance were minimal. Such changes in response format can cause nontrivial differences in response behavior, with differences likely even greater for patient groups who have motor or perceptual difficulties, such as in Parkinson's disease (Begnum & Begnum, 2012).

And yet, the fact that these format changes are major considerations for interpreting cognitive test scores is not always well recognized. A recent prominent example of

a failure to account for such changes was the translation of the NIH Toolbox Cognition tests (Weintraub et al., 2013). The NIH Toolbox is a suite of brief assessments of cognition, sensory, motor and emotional functioning that was developed in 2004 to provide a standard for behavioral assessment in large research cohorts, funded by the large-scale NIH Blueprint for Neuroscience Initiative, in consultation with nearly 300 scientists across 100 academic institutions. In August 2015, the NIH Toolbox was adapted from computerized (laptop/desktop) format to administration on the iPad, in order to accommodate infrastructural changes and the needs of clinical researchers. New tests were released in conjunction with original norms and before the validity of these norms for new tests formats was established. More than a year later, in fall 2016, it was discovered that the norms from the original computerized versions of five of the Cognition assessments were not appropriate for the iPad version. Corrections to norms were released two months later, specific to the iPad version. This example starkly illustrates how—even for the most prominent and sophisticated efforts at digital neuropsychological assessment—an underappreciation of the necessary considerations of digital neuropsychological assessment can substantially threaten the validity of scientific and clinical interpretation (Kane & Parsons, 2017).

In summary, digital neuropsychology needs to pay close attention to differences in response format and stimulus presentation format that vary across digital devices even for tests that rely on the same software, which can differ substantially across modalities. On the positive side, at any given time, there are a fairly limited number of broad formats for digital neuropsychology that can be mostly divided based on screen size (desktop/laptop vs. tablet vs. smartphone) and input type (keyboard, external mouse, trackpad and touchscreen) that still renders a relatively small number of existing combinations that is at least tractable. More investment in validation and collection of normative data is needed when tests are translated across device classes, so that these changes do not threaten the validity of digital neuropsychological assessment.

Challenge 2: device characteristics can introduce systematic measurement bias

In addition to variation between broad device categories (smartphone and tablet), another major barrier to digital neuropsychology lies in the variation between devices and people's relationships with those devices. Digital devices vary in the precision and accuracy of their measurement, based on technical factors such as hardware, software and CPU usage that are beyond the control of the clinician/researcher or patient/participant. Unlike traditional paper-and-pencil tests that are fairly uniform across sites, digital devices vary based on factors mostly determined by the manufacturer of the hardware and software. Such factors are typically not disseminated (or potentially even documented) by hardware and software manufacturers and introduce what are often very difficult to quantify variations in the assessment of behavior.

Differences in the sampling rate of mouse movement or touch, frame rate for dynamic visual displays, or differences in screen size, resolution or the rendering of visual display elements have the potential to influence test performance and the

measurement of behavior. One of the biggest and most underappreciated differences between digital devices is how long it takes different devices to register a user response. Such latency can be operationalized as the total time between when a user makes a response (e.g. taps the screen, clicks the mouse button and presses a key on the keyboard) and when the response is registered by the device. While this latency can be reduced by improvements in hardware and software, it can never be eliminated completely. The measured response time for any test or test trial is the user's true response time plus the device-related response time latency. Unlike desktops and laptops where keyboard and mouse differences in latencies tend to be fairly similar across the same class of input device, there are substantial and often unknown variations in response time latency across touchscreen devices based on differences in hardware and differences in software (Koudritzky, 2016; Koudritzky et al., 2017; Ng, Lepinski, Wigdor, Sanders, & Dietz, 2012; Yun, He, & Zhong, 2017). Technical reports indicate that such touch latencies can differ by as much as 100 ms between popular devices (Siegal, 2013a, 2013b). In the context of tests where the average reaction time is only 200–300 ms (e.g. simple reaction time tests), this represents a substantial proportion of the potential variance (Schatz, Ybarra, & Leitner, 2015). This latency can be objectively measured on a particular device at a particular time using devices such as Google's WALT (Koudritzky et al., 2017), but such objective adjustments are only practical if the landscape of potential devices is fairly limited. While such an assumption might hold for broad device classes (Challenge 1), it does not hold for combinations of hardware and software within those classes. In 2015, there were over 24,000 distinct Android smartphone or tablet devices—each with different potential hardware characteristics—running one of 15+ versions of the Android operating system (OpenSignal, 2015). A small additional investment in validation or collection of normative data would not be sufficient to address this level of technical variability.

But does device variability challenge validity? In the case where device types are randomly distributed across the population—i.e. device familiarity and device ownership are uncorrelated among individuals with similar sociodemographic or clinical characteristics—the increase in noise or random error from digital devices might be acceptable in certain contexts (e.g. large research studies). Unfortunately such variability is *not random*. Ownership of specific digital devices and technologies are related to the same variables that predict neuropsychological performance on reaction-time tests (e.g. education, age, health) (Pew Research Center, 2018; Desilver, 2013). Put another way, device variability can significantly confound the relationship between consequential population and clinical variables and cognitive performance for measures that rely on response time latencies.

In more concrete terms, what this means is that (1) a patient might appear to be impaired due to the effect of being tested on a device with longer latencies, (2) group differences in performance that look like cognitive differences might instead be attributable to differences in device characteristics between two groups (e.g. use of older vs. newer tablets) and (3) an individual measured at two time points on different devices (e.g. by different clinicians) might look like their performance is changing over time, when in fact their neuropsychological functioning has been stable.

Understanding how to deal with such device variability, and the variability in device ownership across the population, will be a critical step in the growth of digital neuropsychology, particularly for settings and study designs that rely on a range of devices. But what is the best solution?

One device to assess them all?

The proposed solution of many cognitive testing developers is to require users to all use the same device hardware and software—and although device variability will still exist, this solution can at least reduce the variation between devices from 100 to 20 ms. Critical software updates or hardware upgrades make this solution perhaps more logistically complicated in practice than it might appear to on the surface, but for certain contexts, mandating use of a single hardware/software combination may, in fact, be the best solution. This is particularly true for certain types of intervention trials, where highly controlled measurement over relatively short time frames is the goal.

But is this a good general solution? The answer depends largely on the context. Consider the fact that efforts to reduce the impact of device variability will necessarily *increase* the impact of device familiarity—if the adopted device is closer to what a person is already familiar with using, they will do better than if the device is not similar to what they already use (McWilliams, Reimer, Mehler, Dobres, & Coughlin, 2015). The impact of device familiarity, in particular, is currently not well understood and warrants further study to characterize the effects of such variability on test performance, both across device classes (smartphones, tablets, laptops/desktops) and within device classes (different types of tablets). This issue has been considered in general terms for the use of computerized assessment, but needs to be revisited now that the landscape of digital technologies has grown so tremendously over the last several years. The trade-off between minimizing the contribution of device variability and device familiarity creates a conundrum for the neuropsychologist: which barrier is more acceptable? Is it better to accept some device-related variability in performance in order to use the devices that the patient or participant is most familiar and comfortable with, or to accept device familiarity differences that will likely inflate scores among individuals who are most familiar with a selected technology?

A single device approach can also be impractical and extremely costly for studies and clinical contexts where a particular measurement is being deployed at scale and over many years. The necessity of buying devices in order to use specific testing software—and potentially different devices for different tests—results in less flexibility and choice for clinicians and researchers, who might choose tests based on existing device resources rather than the most appropriate or valid test for a particular application. Indeed, pragmatic clinical trials (Byrom et al., 2017; Gwaltney et al., 2015) and large cohort studies (NIH, 2018) are increasingly shifting towards a "Bring Your Own Device" (BYOD) model as this reduces costs, participant burden, as well as the potential for user error due to borrowing an unfamiliar device (Armstrong, Semple, & Coyte, 2014; Byrom et al., 2017; McWilliams et al., 2015). In clinical settings, potential applications of digital neuropsychology for patient monitoring (Armstrong et al., 2014) or

screening of patients in remote settings to address global health (Estai et al., 2017; Gomes et al., 2017; Kassianos, Emery, Murchie, & Walter, 2015) will rely on the use of many different devices potentially producing very different data.

Challenge 3: the landscape of digital technology is constantly changing

The final challenge that we articulate here lies in the pace of technology development, itself: the landscape of digital devices and our relationship with those devices changes rapidly and unpredictably.

First, there is the basic fact that devices are rapidly evolving. Each new piece of hardware that is released by a major manufacturer has a new set of characteristics that affect both performance and the user interface. With each operating system or software update, there are potential modifications that could interfere with stimulus presentation, as well as the precision and accuracy of behavioral measurement.

Consider, for example, the touch latency issues described earlier. Device manufacturers are keen to reduce such latencies, which can interfere with the user experience and the practicality of such devices for applications such as drawing, writing and gaming. As a result of this drive, touch latencies have been slowly improving over time (and therefore expected scores), particularly for devices that are marketed to gamers and graphic designers (Yun et al., 2017). The negative consequence of this change is that test scores will start to improve over time due to improvements in technology, with bigger gains for some devices than others. Normative data from devices made by a particular manufacturer (e.g. the iPad/NIH Toolbox) will also become out-of-date or require adjustment relatively quickly. The positive consequence is that as improvements in touchscreen technology shift from exponential to incremental, devices will likely become increasingly similar to each other in characteristics that were previously highly variable.

Yet, even as current technologies improve and become more homogeneous, new technologies are entering the market all the time—and the time it takes from when a new technology goes from cutting edge to standard is daunting for mobile application developers. Consider that the first modern multimedia smartphone, marketed for broad consumer use, was released in 2007 (the iPhone). By 2017, more than three-fourths of adults in the US owned a smartphone. The iPad—the first modern tablet computer—was released in 2010. Now 53% of US adults own a tablet computer. Over the same time period (2010–2018), the percentage of adults who owned a desktop or laptop was stable or slightly dropped (78–73%) (Pew Research Center, 2018). What this means is that, within 10 years, new technologies can rapidly become standards in terms of availability and patient/participant familiarity. The device that is the most familiar for one group might be completely different now than it was 5 years ago. Take for example the observation that tablet computers are easier to learn to use for older adults than traditional desktop computers (Chan, Haber, Drew, & Park, 2016): the motor demands of touchscreens are more intuitive and lower than mouse and keyboard input, and easy-to-use pinch and zoom capabilities allow relatively frictionless accommodation of vision difficulties. What this means is that many older adults may

skip adoption of traditional desktop and laptop computers entirely in favor of tablet computers (Anderson & Perrin, 2017; Tsai, Shillair, Cotten, Winstead, & Yost, 2015). The gap between novelty and ubiquity is now a matter of less than a decade—a pace that we are not used to accommodating in the development of neuropsychological tests.

Proposed solutions

The pursuit of better

We have articulated a set of fairly broad and ubiquitous challenges for digital neuropsychology. Despite these challenges, however, we believe there is cause for optimism. One of the reasons that we are able to identify so many systematic sources of variation in digital technology-based assessment is the fact that digital technology enables the measurement of behavior with enough precision and standardization that these sources of variation become observable. Such observations are difficult or impossible where exogenous sources of variability (i.e. variability not due to differences in neuropsychological functioning) arise from differences in the skill, training and current cognitive status of test administrators (Overton, Pihlsgård, & Elmståhl, 2016), who are fundamental to the timing and precision of measurement in traditional clinical neuropsychology. Indeed, digital neuropsychology shifts many of these administrator sources of variance to the administration device itself—a different problem that requires different solutions.

Here, we attempt to lay out some guidelines for how digital neuropsychology might address or overcome some of the barriers we have described, to preserve the validity of neuropsychological assessment and facilitate innovation. We offer these solutions based on the general understanding that no solution will yield perfect accuracy and precision of measurement—there will always be confounds and potential sources of imprecision that must be understood and dealt with. Nevertheless, attention to potentially addressable issues in digital neuropsychology should facilitate the development and validation of measures that *improve upon the current standard along dimensions that are relevant to a particular clinical or research application*. These might include metrics related to reducing costs, improving accessibility, enabling at-home monitoring or enhancing our understanding of specific cognitive mechanisms.

Solution 1: consider device variability in norming and test design

Although factors related to device variability cannot be eliminated altogether, they can be minimized by adequate attention to norms and designing tests—wherever possible—that are more robust to device variability.

In the same way that classic neuropsychological assessments are designed with clinician or administrator variability in mind (standardized instructions, forms, scoring and training), digital neuropsychological assessments should be designed to minimize the influence of device variability. From a test validation and normative standpoint, this means explicitly testing whether a particular measure produces similar scores across a range of device types. If scores do vary by device, the degree of variability in test scores that is potentially attributable to device type must be quantified. From a test design standpoint, this means reducing reliance on stimuli or measurement characteristics that

are extremely sensitive to device related confounds. More specifically, stimuli should be designed to accommodate a variety of display types and one of the following ways of capturing behavior: (a) where possible, reliance on scores based on accuracy rather than response time; (b) measures with longer reaction times (e.g. average 2000 ms or higher), where variance in response times due to endogenous, neuropsychological factors will typically far exceed variance due to device types (>95%) or (c) measures where scores are calculated with respect to an individual's own baseline on the same or another measure administered using the same device, using subtraction of regression (Munoz, Sliwinski, Scott, & Hofer, 2015; Sliwinski et al., 2018). In the latter case, controlling for scores on another measure with similar device-related motor and perceptual confounds will capture such device-related sources of variability. Such subtraction or regression-based methods are widely used in fields as cognitive neuroscience for quantifying behavior (DeGutis, Wilmer, Mercado, & Cohan, 2013; Mogg, Holmes, Garner, & Bradley, 2008; Redick & Engle, 2006; Susilo, Germine, & Duchaine, 2013), although it should be noted that these scores capture *different information* than scores calculated based on uncorrected reaction times (Lee & Chabris, 2013).

Where reliance on norms across device types is key to test score interpretation, such norms must be updated at a frequency that maps onto the speed of technology development. In this model, the development and updating of norms is a semicontinuous process where new device-specific norms are generated as new devices or major device updates are released, especially if those updates are expected to affect more than 5% of potential users.

Solution 2: pay more attention to user interface design

Differences in user technical experience and device familiarity may interfere with valid neuropsychological assessment using digital tools, and therefore, user variability must be a primary consideration in digital neuropsychology. This means explicit attention to developing user interfaces that are accessible and engaging across a range of devices, sociodemographic groups and expected clinical characteristics. Decades of human computer interaction research have demonstrated that the way software is designed changes the behavior that is elicited from the user (Norman & Draper, 1986). One of the things that catapulted the smartphone and tablet into widespread ubiquity across a broad consumer base was Apple's attention to design simple user interfaces, essentially creating a new market of users who were able to pick up devices and immediately understand how to interact with that device with minimal technical expertise. The fundamental principles of simplicity and clarity can be applied to neuropsychological assessments. For example, people are not very good at reading and retaining written instructions. Structured examples and practice trials that *teach the user how to interact with the test* are far more useful and virtually essential when transitioning to digital neuropsychological assessments (Johnson, 2013). In addition, tasks that help a user become more familiar with a particular device can be a helpful way of reducing potential anxiety or difficulty with new technologies. This has been done for virtual reality-based neuropsychological tests where user familiarity with the technology tool is expected to be low (e.g. AULA; Climent & Banterla, 2011).

Solution 3: treat tests as software

Third, and most importantly, is the fundamental understanding that in digital neuropsychology, tests are software. This means that all of the best practices for user-centered software development need to be integrated into the design, development and lifecycle of digital neuropsychological assessments (Abrahamsson, Salo, Ronkainen, & Warsta, 2017; Brhel, Meth, Maedche, & Werder, 2015). For example, digital neuropsychological assessments will need regular updates that allow the test to be modified to accommodate changes in technology and ensure continued technical compatibility and usability; robust systems for version control that include identification of features, bug fixes and modifications that might impact software compatibility, performance and behavioral measurement; and incorporation of best practices for developing and evaluating user interfaces that might—in some cases—need to be customized for a target population (Begnum & Begnum, 2012). As discussed in the case of norms, a software development approach to digital neuropsychology and innovation relies on methods that are continuous and iterative. The reliance on static metrics with minimal changes in format over many years (a model adapted from print publishing) does not work in a digital context as it fails to consider both the technological and social context in which that software must operate. Instead, neuropsychological tests must be viewed as continual works-in-progress: pieces of software that are continually refined, fine-tuned and validated through cycles of modification and evaluation before production-ready versions of that software are released.

Opportunities for innovation

The most obvious benefits of digital neuropsychology will be to reduce the cost of neuropsychological assessment and increase the accessibility of neuropsychological services, particularly in rural or low-income populations (Kane & Parsons, 2017). In our view, however, the most transformative opportunity digital neuropsychology can offer is the ability to conduct frequent ambulatory neuropsychological assessments in a person's everyday environment (Sliwinski et al., 2018). Neuropsychological assessment is typically limited by pure logistics: assessment takes place in a clinic or laboratory at one (or, at best, a few) time points. Shorter and more frequent assessment outside of the clinic can enable (1) more reliable estimates of neuropsychological functioning averaged across time, (2) the ability to capture variability of a person's neuropsychological functioning over time and (3) better ecological validity. Automated collection of cognitive data can then by combined with relevant passive sensor data to improve interpretation and enable machine learning approaches for real-time detection of risk (Cook, Schmitter-Edgecombe, & Jonsson, 2018).

Better estimates of typical neuropsychological functioning

Most neuropsychological assessment relies on estimates of performance at a single point in time or at two to three time points separated by a long interval (e.g. months and years). Yet, we know that neuropsychological functioning can vary over time, based on many factors including sleep quality (Gamaldo, Allaire, & Whitfield, 2010),

time-of-day (Riley, Esterman, Fortenbaugh, & DeGutis, 2017), stress (Hyun, Sliwinski, & Smyth, 2018), glycemic status (Gold, MacLeod, Deary, & Frier, 1995) and physical activity (Brisswalter, Collardeau, & René, 2002). Estimates of neuropsychological functioning are thus dependent on a variety of state variables that may or may not represent the typical conditions of a person's everyday life (Arnett, 2013). This introduces *temporal sampling error* that reduces the reliability of neuropsychological assessments, interfering with our ability to diagnose underlying brain disease, predict individual outcomes, or evaluate the impact of interventions (Sliwinski, Almeida, Smyth, & Stawski, 2009). Methods for ambulatory assessment that rely on shorter but more frequent assessments using digital devices can significantly address such error and have been shown to improve sensitivity for detecting change over time in longitudinal studies (Sliwinski et al., 2018). The additional information provided by frequent ambulatory assessments, performed between traditional clinic-based assessments, could dramatically improve detection of brain dysfunction, as well as enable more precision in determining conversion from prodromal states (e.g. mild cognitive impairment) to dementia and recovery from acute brain injury (e.g. return to baseline after mild traumatic brain injury).

Variability as neuropsychological indicator

In addition to permitting better estimates of average neuropsychological functioning, the same ambulatory assessment methods can enable estimates of variability of neuropsychological functioning over time (days, weeks and months) in individuals, which can be an important indicator of brain health. Greater variability in neuropsychological performance has been linked with changes in brain health associated with age (Hultsch & MacDonald, 2004), traumatic brain injury (Cole, Gregory, Arrieux, & Haran, 2018), epilepsy (Srnka, Seidenberg, Hermann, & Jones, 2018) and dementia (Holtzer, Verghese, Wang, Hall, & Lipton, 2008). Variability in neuropsychological test scores over time may provide new information for evaluating clinical status and prediction of outcomes that is not currently captured by measures based on single time point or average performance (Sliwinski et al., 2018).

Ecological validity and context

The goal of traditional neuropsychological assessment is to estimate an individual's *best* performance (Chaytor & Schmitter-Edgecombe, 2003; Long & Collins, 1997). Such assessments are useful for determining an individual's potential capabilities but do not necessarily reflect how a person actually functions in their natural environment. Clinically, it is common for a patient to perform normally on neuropsychological testing in a controlled office environment, but report cognitive problems in daily life. Clinicians often have to rely solely on the subjective reports of patients to determine what environmental factors may be adversely impacting cognitive performance. Self-report is often biased in systematic ways, particularly for individuals with brain dysfunction. To understand everyday neuropsychological functioning and make accurate predictions about a person's ability to work, attend school and participate in other activities, it is necessary to assess people in everyday contexts (Chaytor & Schmitter-

Edgecombe, 2003). The downside of assessment in naturalistic environments is that it is hard to know *what* factors might be contributing to poorer performance or to what degree—contextual factors such as background noise, physical location and environmental distractions can make it harder to interpret test scores, particularly when such factors are unmeasured or unknown. Today, personal digital devices are increasingly equipped with a range of sensors that make it possible to measure contextual factors such as recent activity, GPS location, ambient noise levels and environmental distractions, as well as momentary assessment of emotional state, in a way that can allow us to interpret neuropsychological functioning with respect to important contextual factors (Cook, Schmitter-Edgecombe, & Jonsson, 2018). While significant validation work still remains to be done to make such digital sensor data useful in everyday clinical care, they provide a promising avenue of exploration for helping us to understand variations in neuropsychological functioning that are related to everyday environmental factors and ways to maximize cognitive performance in daily life.

Concluding thoughts: toward open neuropsychology

Here, we argue that digital neuropsychology requires a fundamental shift in the way we conceptualize development and innovation in neuropsychology. Our hope is that the guidelines and opportunities discussed here will aid the field in keeping all that is good about traditional clinical neuropsychology, while taking advantage of new digital approaches in a way that accelerates the pace of innovation without threatening validity.

As in many areas of measurement, we believe the future of neuropsychology will be digital. Beyond a matter of reduced cost and higher accessibility, such a transition will be critical to the viability of neuropsychology *as a field* which will suffer if it stagnates—and it will stagnate if we do not take advantage of the potential of digital devices to capture and quantify the minds and brains of individuals, especially as we move towards precision medicine approaches in healthcare. As with all things, however, uncritical acceptance of digital neuropsychology as simply a change in format will also limit progress. Instead, we advocate a broader awareness of the opportunities and challenges inherent in digital approaches to neuropsychology, so that we are better positioned to build a future of neuropsychological assessment that is scientifically robust, inclusive and innovates at the pace of digital technology.

An important consequence of a shift to digital neuropsychology will be an increased burden of norms development and dissemination—an issue that is already a limiting factor in traditional neuropsychology. The cost of building and maintaining digital tools and norms for those tools requires some thoughtfulness around long-term sustainability, and this problem has not yet been solved in the public sector. Indeed, the only groups with incentives and resources to develop and validate at the rate required for digital neuropsychology may be commercial test developers. Many neuropsychologists worry that such a heavy reliance on the commercial sphere may further deepen the information asymmetry between test developers and clinicians/researchers, who must evaluate the validity of tests and quality of their norms without access to the data that was used to establish validity.

Organizations dedicated to open source software and related solutions—companies such as the nonprofit Center for Open Science (Foster & Deardorff, 2017; Nosek et al., 2015) and Sage Bionetworks (Bot et al., 2016; Wilbanks & Friend, 2016)—have demonstrated it is possible to build viable business models around open source software in ways that could be translated to digital neuropsychology. The sustainability of such business models in the longer term is still unclear, however but provides promising avenues for long-term development. It will be important—as a community—that we thoughtfully consider how we might build open source measures for digital neuropsychology that enable the continuous and community-based development of norms. One part of this shift will be demanding, at minimum, that commercial testing companies make norms data freely available to enable their evaluation by the research and clinical communities.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2017). Agile software development methods: Review and analysis. *arXiv*. preprint arXiv:1709.08439.

Anderson, M., & Perrin, A. (2017). *Technology use among seniors*. Washington, DC: Pew Research Center.

Armstrong, K. A., Semple, J. L., & Coyte, P. C. (2014). Replacing ambulatory surgical follow-up visits with mobile app home monitoring: Modeling cost-effective scenarios. *Journal of Medical Internet Research*, *16*(9), e213.

Arnett, P. (Ed.). (2013). *Secondary influences on neuropsychological test performance*. Oxford: Oxford University Press.

Bailey, S. K., Neigel, A. R., Dhanani, L. Y., & Sims, V. K. (2017). Establishing measurement equivalence across computer- and paper-based tests of spatial cognition. *Human Factors*, 60, 340–350. 0018720817747731.

Begnum, M. E. N., & Begnum, K. M. (2012). On the usefulness of off-the-shelf computer peripherals for people with Parkinson's Disease. *Universal Access in the Information Society*, *11*(4), 347–357.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., … Trister, A. D. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, *3*, 160011.

Brhel, M., Meth, H., Maedche, A., & Werder, K. (2015). Exploring principles of user-centered agile software development: A literature review. *Information and Software Technology*, *61*, 163–181.

Brisswalter, J., Collardeau, M., & René, A. (2002). Effects of acute physical exercise characteristics on cognitive performance. *Sports Medicine*, *32*(9), 555–566.

Buckley, R. F., Sparks, K. P., Papp, K. V., Dekhtyar, M., Martin, C., Burnham, S., … Rentz, D. M. (2017). Computerized cognitive testing for use in clinical trials: A comparison of the NIH Toolbox and Cogstate C3 batteries. *The Journal of Prevention of Alzheimer's Disease*, *4*(1), 3.

Byrom, B., Muehlhausen, W., Flood, E., Cassedy, C., Skerritt, B., & Mc Carthy, M. (2017). Patient attitudes and acceptability towards using their own mobile device to record patient reported outcomes data in clinical trials. *Scoliosis*, *6*, 4.

Pew Research Center. (2018). *Demographics of mobile devices ownership in the United States*. Retrieved from http://www.pewinternet.org/fact-sheet/mobile/

Chan, M. Y., Haber, S., Drew, L. M., & Park, D. C. (2016). Training older adults to use tablet computers: Does it enhance cognitive function? *The Gerontologist*, *56*(3), 475–484.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychology Review*, *13*(4), 181–197.

Climent, G., & Banterla, F. (2011). *AULA, ecological evaluation of attentional processes*. San Sebastian: Nesplora.

Cook, D. J., Schmitter-Edgecombe, M., & Jonsson, L. (2018). Technology-enabled assessment of functional health. *IEEE Reviews in Biomedical Engineering*.

Cole, W. R., Gregory, E., Arrieux, J. P., & Haran, F. J. (2018). Intraindividual cognitive variability: An examination of ANAM4 TBI-MIL simple reaction time data from service members with and without mild traumatic brain injury. *Journal of the International Neuropsychological Society*, *24*(2), 156–162.

Corrigan, J. D., & Hinkeldey, N. S. (1987). Relationships between parts A and B of the trail making test. *Journal of Clinical Psychology*, *43*(4), 402–409.

Crowe, S. F. (1998). The differential contribution of mental tracking, cognitive flexibility, visual search, and motor speed to performance on parts A and B of the trail making test. *Journal of Clinical Psychology*, *54*(5), 585–591.

Daniel, M. H., Wahlstrom, D., & Zhang, O. (2014). *Equivalence of Q-interactive™ and paper administrations of cognitive tasks: WISC®–V* (Q-interactive Technical Report 8).

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

DeGutis, J., Wilmer, J., Mercado, R. J., & Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, *126*(1), 87–100.

Desilver, D. (2013). *As it turns 6, a look at who uses the iPhone (no, not 'everybody')*. Washington, DC: Pew Research Center.

Díaz-Orueta, U., Garcia-López, C., Crespo-Eguílaz, N., Sánchez-Carpintero, R., Climent, G., & Narbona, J. (2014). AULA virtual reality test as an attention measure: Convergent validity with Conners' Continuous Performance Test. *Child Neuropsychology*, *20*(3), 328–342.

Estai, M., Kanagasingam, Y., Xiao, D., Vignarajan, J., Bunt, S., Kruger, E., & Tennant, M. (2017). End-user acceptance of a cloud-based teledentistry system and android phone app for remote screening for oral diseases. *Journal of Telemedicine and Telecare*, *23*(1), 44–52.

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association*, *105*(2), 203.

Gamaldo, A. A., Allaire, J. C., & Whitfield, K. E. (2010). Exploring the within-person coupling of sleep and cognition in older African Americans. *Psychology and Aging*, *25*(4), 851.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857.

Giannouli, E., Bock, O., & Zijlstra, W. (2018). Cognitive functioning is more closely related to real-life mobility than to laboratory-based mobility parameters. *European Journal of Ageing*, *15*(1), 57–65.

Gold, A. E., MacLeod, K. M., Deary, I. J., & Frier, B. M. (1995). Hypoglycemia-induced cognitive dysfunction in diabetes mellitus: Effect of hypoglycemia unawareness. *Physiology & Behavior*, *58*(3), 501–511.

Gomes, M. S., Bonan, P. R. F., Ferreira, V. Y. N., de Lucena Pereira, L., Correia, R. J. C., da Silva Teixeira, H. B., … Bonan, P. (2017). Development of a mobile application for oral cancer screening. *Technology and Health Care*, *25*(2), 187–195.

Gwaltney, C., Coons, S. J., O'Donohoe, P., O'Gorman, H., Denomey, M., Howry, C., & Ross, J. (2015). "Bring your own device" (BYOD): The future of field-based patient-reported outcome data collection in clinical trials? *Therapeutic Innovation & Regulatory Science*, *49*(6), 783–791.

Holtzer, R., Verghese, J., Wang, C., Hall, C. B., & Lipton, R. B. (2008). Within-person across-neuropsychological test variability and incident dementia. *The Journal of American Medical Association*, *300*(7), 823–830.

Hultsch, D. F., & MacDonald, S. W. (2004). Intraindividual variability in performance as a theoretical window onto cognitive aging. *New Frontiers in Cognitive Aging*, *65*, 88.

Hyun, J., Sliwinski, M. J., & Smyth, J. M. (2018). Waking up on the wrong side of the bed: The effects of stress anticipation on working memory in daily life. *The Journal of Gerontology: Series B*.

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. *The Journal of American Medical Association*, *318*(13), 1215–1216.

Iriarte, Y., Diaz-Orueta, U., Cueto, E., Irazustabarrena, P., Banterla, F., & Climent, G. (2016). AULA—Advanced virtual reality tool for the assessment of attention: Normative study in Spain. *Journal of Attention Disorders*, *20*(6), 542–568.

Johnson, J. (2013). *Designing with the mind in mind: Simple guide to understanding user interface design guidelines*. Amsterdam: Elsevier.

Kane, R. L., & Parsons, T. D. (2017). *The role of technology in clinical neuropsychology*. Oxford: Oxford University Press.

Kassianos, A., Emery, J., Murchie, P., & Walter, F. M. (2015). Smartphone applications for melanoma detection by community, patient and generalist clinician users: A review. *British Journal of Dermatology*, *172*(6), 1507–1518.

Koudritzky, M. (2016). A new method to measure touch and audio latency. Retrieved from https://android-developers.googleblog.com/2016/04/a-new-method-to-measure-touch-and-audio.html

Koudritzky, M., Fair, B., Jain, S., Quinn, P., Turner, D., Frysinger, M., & Wimmer, R. (2017). WALT Latency Timer. *GitHub*.

Lee, J. J., & Chabris, C. F. (2013). General cognitive ability and the psychological refractory period: Individual differences in the mind's bottleneck. *Psychological Science*, *24*(7), 1226–1233.

Long, C. J., & Collins, L. F. (1997). Ecological validity and forensic neuropsychological assessment. In *The practice of forensic neuropsychology: Meeting challenges in the courtroom* (pp. 153–164). New York: Plenum Press.

McWilliams, T., Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2015). *Effects of age and smartphone experience on driver behavior during address entry: A comparison between a Samsung Galaxy and Apple iPhone*. Paper presented at the proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications.

Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., & Hong, J. I. (2014). *Toss'n'turn: Smartphone as sleep and sleep quality detector*. Paper presented at the proceedings of the SIGCHI conference on human factors in computing systems.

Mogg, K., Holmes, A., Garner, M., & Bradley, B. P. (2008). Effects of threat cues on attentional shifting, disengagement and response slowing in anxious individuals. *Behaviour Research and Therapy*, *46*(5), 656–667.

Munoz, E., Sliwinski, M. J., Scott, S. B., & Hofer, S. (2015). Global perceived stress predicts cognitive change among older adults. *Psychology and Aging*, *30*(3), 487–499.

National Institutes of Health, All of Us Research Program. (2018). All of Us Research Program: Operational Protocol. Retrieved from https://allofus.nih.gov/sites/default/files/aou_operational_protocol_v1.7_mar_2018.pdf

Ng, A., Lepinski, J., Wigdor, D., Sanders, S., & Dietz, P. (2012). *Designing for low-latency direct-touch input*. Paper presented at the proceedings of the 25th annual ACM symposium on user interface software and technology, Cambridge, MA, USA.

Norman, D. A., & Draper, S. W. (1986). *User centered system design: New perspectives on human–computer interaction*. Boca Raton, FL: CRC Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science (New York, NY)*, *348*(6242), 1422–1425.

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, *41*(7), 1691.

OpenSignal. (2015). *Android fragmentation visualized*. Retrieved from https://opensignal.com/legacy-assets/pdf/reports/2015_08_fragmentation_report.pdf

Overton, M., Pihlsgård, M., & Elmståhl, S. (2016). Test administrator effects on cognitive performance in a longitudinal study of ageing. *Cogent Psychology*, *3*(1), 1260237.

Redick, T. S., & Engle, R. W. (2006). Working memory capacity and attention network test performance. *Applied Cognitive Psychology*, *20*(5), 713–721.

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327.

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276.

Riley, E., Esterman, M., Fortenbaugh, F. C., & DeGutis, J. (2017). Time-of-day variation in sustained attentional control. *Chronobiology International*, *34*(7), 993–1001.

Schatz, P., Ybarra, V., & Leitner, D. (2015). Validating the accuracy of reaction time assessment on computer-based tablet devices. *Assessment*, *22*(4), 405–410.

Siegal, J. (2013a). Here's why typing on Android phones is harder than typing on an iPhone. *BGR*. Retrieved from bgr.com website: http://bgr.com/2013/09/20/iphone-android-touchscreen-responsiveness/

Siegal, J. (2013b). Study: iPads are the most responsive tablets in the world. Retrieved from http://bgr.com/2013/10/09/tablet-touch-screen-responsiveness/

Sliwinski, M. K., Almeida, D. M., Smyth, J., & Stawski, R. S. (2009). Intraindividual change and variability in daily stress processes: Findings from two measurement burst studies. *Psychology and Aging*, *24*(4), 828.

Sliwinski, M. J., Mogle, J. A., Hyun, J., Munoz, E., Smyth, J. M., & Lipton, R. B. (2018). Reliability and validity of ambulatory cognitive assessments. *Assessment*, *25*(1), 14–30.

Srnka, K., Seidenberg, M., Hermann, B., & Jones, J. (2018). Intraindividual variability in attentional vigilance in children with epilepsy. *Epilepsy & Behavior*, *79*, 42–45.

Susilo, T., Germine, L., & Duchaine, B. (2013). Face recognition ability matures late: Evidence from individual differences in young adults. *Journal of Experimental Psychology: Human Perception & Performance*, *39*(5), 1212–1217.

Tsai, H-y. S., Shillair, R., Cotten, S. R., Winstead, V., & Yost, E. (2015). Getting grandma online: Are tablets the answer for increasing digital inclusion for older adults in the US? *Educational Gerontology*, *41*(10), 695–709.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S54–S64.

Wilbanks, J., & Friend, S. H. (2016). First, design for data sharing. *Nature Biotechnology*, *34*(4), 377.

Woodward, J., Shaw, A., Luc, A., Craig, B., Das, J., Hall, P., Jr, … Brown, Q. (2016). *Characterizing how interface complexity affects children's touchscreen interactions*. Paper presented at the proceedings of the 2016 CHI conference on human factors in computing systems.

Yun, M. H., He, S., & Zhong, L. (2017). *Reducing latency by eliminating synchrony*. Paper presented at the proceedings of the 26th international conference on world wide web, Perth, Australia.