The Technology Crisis in Neuropsychology

Justin B. Miller1,*, William B. Barr2

1Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, USA
2NYU, School of Medicine, New York, NY, USA

*Corresponding author at: Cleveland Clinic Lou Ruvo Center for Brain Health, 888 W. Bonneville Ave, Las Vegas, NV 89106, USA.
Tel.: +1-702-483-6000; fax: +1-702-483-6039.
E-mail address: millerj4@ccf.org (J.B. Miller).

Abstract

Neuropsychology has fallen reliant on outdated and labor intensive methods of data collection that are slow, highly inefficient, and expensive, and provide relatively data-poor estimates of human behavior despite rapid technological advance in most other fields of medicine. Here we present a brief historical overview of current testing practices in an effort to frame the current crisis, followed by an overview of different settings in which technology can and should be integrated. Potential benefits of laboratory based assessments, remote assessments, as well as passive and high-frequency data collection tools rooted in technology are discussed, along with several relevant examples and how these technologies might be deployed. Broader issues of data security and privacy are discussed, as well as additional considerations to be addressed within each setting. Some of the historical barriers to adoption of technology are also presented, along with a brief discussion of the remaining uncertainties. While by no means intended as a comprehensive review or prescriptive roadmap, our goal is to show that there are a tremendous number of advantages to technologically driven data collection methods, and that technology should be embraced by the field. Our predictions are that the comprehensive assessments of the future will likely entail a combination of lab-based assessments, remote assessments, and passive data capture, and leading the development of these efforts will cement the role of neuropsychology at the forefront of cognitive and behavioral science.

Keywords: Computer science; Cognition; Behavior; Neuroscience; Information technology; Neurology; Assessment

A Brief Historical Review and Statement of the Problem

Based on emerging evidence, it has become increasingly clear that neuropsychology has essentially remained stagnant and reluctant to integrate technology into practice (Rabin et al., 2014). As a result, the field has fallen reliant on outdated and labor intensive methods of data collection that are slow, highly inefficient, and expensive, and provide relatively data-poor estimates of human behavior (Collins & Riley, 2016; Leurent & Ehlers, 2015) in an era of technological growth and development that dwarfs any other time period in history. Our failure to integrate technology directly impedes advancement of cognitive science and the relevance of our field when many other fields of medicine have readily embraced technology, even though integrating technology into cognitive assessment practices has tremendous potential to advance the field at a faster and more continuous pace. The primary aim of the present paper is to preview several potential advantages afforded by greater integration of technology, balanced by some of the most salient considerations that need to be addressed during both development and implementation of any technological advancement. While by no means intended as a comprehensive review or prescriptive roadmap, our goal is to show that there are a tremendous number of advantages to technologically driven data collection methods, and that technology should be embraced by the field.

To establish a context and frame the origins of the current crisis, it is important to begin our discussion with a brief review of the history of clinical neuropsychology and the evolution of its use of technology. Although it is clear that psychologists
and clinicians from other fields were using psychological methods to study brain–behavior relations in animals and humans back in the second half of the 19th Century, the discipline of clinical neuropsychology, as we know it today, originated more recently with the establishment of specialty journals and professional organizations in the 1960s and 1970s (Meier, 1992). During that early time period, there were changes in the field, characterized in a recent review as a transformation from Neuropsychology 1.0 to 2.0 (Bilder, 2011), whereby clinicians evolved from the use of individual measures of brain functioning with limited norms to the use of tests validated through more sophisticated psychometric and normative approaches.

Looking at the tests that have been used by neuropsychologists over the years, one finds that few of them were developed solely for the assessment of brain–behavior relationships. As many know, the history of psychological testing began with the initial adaptation of instruments developed in the laboratory for the study of individual differences, which continued through development of measures designed to address needs arising in various applied settings, such as classifying students, soldiers, and immigrants (Geisinger, 2000). A listing of those settings and others relevant to the development of neuropsychological testing is provided in Table 1.

What one finds is that many of the tests used in our field for assessment of brain functions has resulted from the adaptation of pre-existing measures originally developed for other purposes. Nowhere is this more evident than in the development of the Boston Process Approach (Kaplan, 1988), where the limitations inherent to existing tests led to modifications and expansion of those measures in a prescribed manner to capture qualitative features of the test performance in patients with a range of neurological, psychiatric, and medical conditions.

Although many other fields in healthcare and neuroscience have made significant technological advances over the years, neuropsychological testing continues to rely primarily on paper-and-pencil tasks. This has prolonged neuropsychology’s tie to traditional test publishing companies, who by nature are more focused on maintaining revenue through sales of test forms, stimulus books, and manuals than making any massive changes in their products to keep up with ongoing scientific and technological developments. Tracing the history of test publishing, it is interesting to note that psychological tests were initially produced and distributed by psychologists, starting with James McKeen Cattell and his development of The Psychological Corporation in the 1920s (Sokal, 1981). The Psychological Corporation was eventually acquired by a larger global publishing company, in which distribution of psychological tests accounts for only a fraction of annual sales exceeding 5-billion-dollars.

Publishers of these analog tests have made at least minor efforts to update the tests sold currently for use in neuropsychological assessment batteries. However, many continue to sell minimally revised versions of older tests, using the same stimuli and materials that were developed in non-clinical settings more than 100 years ago. For example, among the stimuli sold in the current versions of the Wechsler Adult Intelligence Scales (e.g., the Fourth Edition; WAIS-IV; Wechsler, 2008), one can find many modernized versions of the same pictures and materials from tests developed in the 1880s through World War I (Boake, 2002).

Fig. 1 presents an example of an item included in the Picture Completion subtest, which has been adapted for use in various versions of the of the Wechsler Intelligence Scales. The figure originated in the Group Examination Beta, a test designed to assess intelligence in military recruits during World War I (Yerkes, 1921). A modernized version of the same figure remains in the current edition of the WAIS-IV (Wechsler, 2008), depicted as a brown-skinned man combing his hair in a mirror. Similarly, the methodology commonly used today in assessment of verbal learning was developed back in 1885 when association-theory was the dominant model of cognitive functioning. Some of the list-learning tests we use today (e.g., Rey Auditory Verbal Learning Test) even continue to use the same word lists that were developed back in 1919 (Boake, 2002; Ebbinghaus, 1913).

Many believed that neuropsychological testing was about to undergo a significant paradigm shift following the large-scale sales and distribution of low-cost microcomputers to scientists and the public in the 1980s. There was, indeed, a small faction in the neuropsychological community who believed that computer-assisted test administration would provide the field with increased control over test administration, scoring, and data collection (Kane & Kay, 1992). However, rather than focusing on the many advantages inherent to computer testing, most of those practicing in the field chose to focus on the technological challenges associated with computer settings (Cernich, Brennana, Barker, & Bleiberg, 2007), concerns about clinicians and

<table>
<thead>
<tr>
<th>Source</th>
<th>Key figures</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Psychology</td>
<td>Cattell and Franz</td>
<td>1880s–1910s</td>
</tr>
<tr>
<td>French Educational System</td>
<td>Binet and Simon</td>
<td>1900s–1930s</td>
</tr>
<tr>
<td>United States Immigration</td>
<td>Knox</td>
<td>1910s</td>
</tr>
<tr>
<td>United States Army Classification</td>
<td>Yerkes and Otis</td>
<td>1910s</td>
</tr>
<tr>
<td>Clinical Psychology</td>
<td>Rappaport, Gill, and Shafer</td>
<td>1940s–1950s</td>
</tr>
<tr>
<td>Neurology/Neurosurgery</td>
<td>Benton, Teuber, Halstead, Reitan, and Luria</td>
<td>1940s–1970s</td>
</tr>
</tbody>
</table>
subjects lacking familiarity with the new technology, and the possibility of a diminishing role for examiners who are trained in evaluation and interpretation of the test results (and many of these same concerns persist today, as will be discussed).

Another likely factor contributing to the limited adoption of technology by many neuropsychologists at the time was the relative lack of innovation in the tests that were offered for use in the computer administration. Rather than incorporating many of the advances in cognitive neuroscience or computer science, most test developers chose simply to redevelop paper-pencil methods, such as measures of digit-span or coding, for administration on the computer. In most cases, the requisite psychometric studies, examining concurrent validity of the computer measures in relation to the paper-pencil standards were not performed, which limited adoption even further. The quantity and quality of the normative data obtained for the computer measures also often paled in comparison to what had been available for their analog counterparts.

While sales and distribution of a limited number of computerized tests were successful (e.g., the Continuous Performance Task; Beck, Bransome, Mirsky, Rosvold, & Sarason, 1956), test publishers remained focused on developing electronic scoring applications for paper-pencil tests, allowing them to continue to produce and sell the analog tests. In response, the development and sales of computerized test batteries was taken up by smaller companies who proceeded to invest more in the technology itself rather than in the psychometric properties of the tests or the quality of the normative data. In many cases, these test batteries were marketed directly to clinicians practicing outside the neuropsychological community, with a focus on assessing frequent and high profile conditions, such as ADHD or concussion. Unfortunately, studies performed on these tests following sales to the public have found that many of them demonstrate insufficient psychometric properties for clinical use (Fratti, Bowden, & Cook, 2017; Resch, McCrea, & Cullum, 2013). As a result, many neuropsychologists remain skeptical about the use of these computer test batteries and have been hesitant to incorporate them in their practice.

It might be easy to express views and make conclusions on the limited degree to which the field of neuropsychology is following technological advances and adapting to new trends, but it is another thing to support those conclusions with data. In the first of a series of practice surveys performed on neuropsychology practitioners over the past several years (Rabin, Barr, & Burton, 2005), it was found that the tests used most frequently by neuropsychologist consisted of standard paper-pencil tests including the Wechsler scales (Intelligence & Memory), Trail Making Test, and California Verbal Learning Test (CVLT-II). A follow-up survey performed 10 years later revealed that neuropsychologists were continuing to use these same tests, with no tangible signs of any evolution in the field or change in clinical practice patterns (Rabin, Paolillo, & Barr, 2016). Looking further into the use of technology, the survey investigators found that, of the 693 different tests listed by respondents in the study, only 6% were computerized versions of standardized tests or computerized test batteries (Rabin et al., 2016). In a separate study, the investigators found that 45% of the neuropsychologists completing the same survey reported never using computer testing in their practice with only 18% reporting that they used computerized methodology in their practice on a frequent basis (Rabin et al., 2014). In explanation of these results, the authors offer financial costs associated with the tests, lack of appropriate normative data, and concerns about test utility or validity as possible reasons for the low rates of computer usage.

The survey findings, taken together, confirm our impressions that the field of neuropsychology is failing to keep up with many other scientific and healthcare professions as a result of its lack of adaption to ongoing advances in technology. What is

![Fig. 1. Item 18 from Picture Completion, Group Examination Beta, Preliminary Form, Test 5 (Yerkes, 1921).](https://academic.oup.com/acn/article-abstract/32/5/541/3852214/543)
evident is that the vast majority of clinicians in the field continue to embrace the use of analog tests based on theoretical concepts and stimuli developed more than 100 years ago. In most cases, these tests were not developed for assessment of brain–behavior relationships, leading neuropsychology practitioners to develop a number of modifications to enable these tests to meet their clinical needs (e.g., Boston Process Approach). While there is no doubt that, as a result of the field’s rigorous education training model, clinical neuropsychologists are the professionals most qualified to assess cognitive and behavioral disturbances associated with brain disorders, the field risks an assault on its integrity in the eyes of the public and other professionals if it continues to lag in its adaptation to the digital age.

Advantages of Technological Integration and How Might We Use It?

Aside from neuropsychology, the rest of the modern world is ubiquitously immersed in technology, including most healthcare sectors. As the underlying technology evolves, individuals are leaving behind increasingly sophisticated digital footprints as a function of everyday behavior. There are a number of ways these digital footprints and related technology can and should be integrated into standard neuropsychology practice and more broadly, the quantification and measurement of human cognition and behavior. From technology driven assessments in the lab or clinic to passive data collection (PDC) form everyday behaviors, each implantation of technology in behavioral measurement has its own advantages. Looking ahead, the most comprehensive profiles of human behavior will likely emerge from an integrated approach that draws from one or more technological resources.

Laboratory Testing

The most straightforward approach is to simply incorporate technology into our current assessment methods and digitize existing tests by porting them to a tablet or computer. Despite the dismal track record of similar efforts to date, this is essentially what several of the major test publishers have moved to, or are moving towards, and soon may be the only option. And although digitizing current tests certainly has some advantages over analog testing, none of these efforts reflect truly novel developments, essentially, turning to computers to automate administration and scoring. Technology could be leveraged far more effectively if, in addition to computer based administration and scoring, development efforts also focused on generating new data points from standard administration of current tests. For example, there are numerous cognitive tasks that require drawing or written responses, and any of these could be easily done on a tablet or using a “smart pencil” to record the written output. In addition to being able to record, review, and “replay” responses, there are a tremendous number of additional metrics that could be readily extracted and quantified, including the speed and consistency of output over time, accuracy, pencil pressure, the number, length, and position of pauses, and the amplitude and frequency of tremulousness differentiated between at rest and while moving. Efforts to validate traditional pencil drawing tasks administered on a tablet are already in various stages of development including the Clock Drawing Task (Cohen et al., 2014) and Trail Making Test (Fellows, Dahmen, Cook, & Schmitter-Edgecombe, 2016), and the future of these and other measures look promising.

Many non-drawing tasks could also be administered via a digital platform. For example, the Tower of Hanoi is easily administered on a tablet and multiple versions already exist for download. Of course, most are variants of the standardized test developed for entertainment purposes and would require standardization prior to clinical use, but the basic principles of the task apply, readily demonstrating feasibility. Converting the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983) or some other version of a two-dimensional picture naming test to a tablet would be incredibly simple (e.g., as a slideshow); however, a more sophisticated naming task that utilizes dynamic three-dimensional digital stimuli that could be manipulated by the test taker (e.g., rotated, zoomed, etc.) or superimposed onto images of real-world context as an alternative to the semantic or phonemic cue could be relatively easily developed. A visuospatial construction task (e.g., using blocks) could be administered using similar digital three-dimensional stimuli, or tangible stimuli could be retained but with a built-in accelerometer to record the number of manipulations prior to placement and internal sensors to automate scoring; such a task is exemplified by the Sensor-integrated Geometric Blocks (SIG-Blocks; Lee, Jeong, Schindler, & Short, 2016). Similar sensor-based technology could be developed for any task with tangible stimuli that are manipulated by test-takers. Verbal fluency tasks could be developed that no longer require manual recording of responses and instead rely on an audio recording and speech recognition technology to automatically record latency, pauses, grammatical errors, consistency, and evaluate patterns of responding (e.g., semantic vs. phonemic clustering) in addition to traditional scoring metrics. Perhaps one of the most obvious points for additional behavioral data is reaction time. Computers have a substantially greater range and degree of precision than humans, and despite individual hardware and software parameters that have historically posed problems (Cernich et al., 2007), these issues could potentially be addressed during development (e.g., developing programs that automatically detect
and adapt to unique hardware configurations) and readily incorporated into essentially every cognitive task. If reaction time were measured over the course of an evaluation and correlated with task demands, it could potentially eliminate the need for a dedicated measure of processing speed.

On the surface, assessing episodic memory seems somewhat less amenable to technology integration in comparison to other cognitive domains. For both verbal and nonverbal tasks, presenting stimuli for learning trials is relatively straightforward, as is assessing recognition memory. Free recall, however, poses a greater challenge. For verbal tasks, speech recognition could be utilized to record responses. These recordings could be saved and played back for later verification, if needed, as there would certainly be limitations and confounds for some individuals that might interfere with speech recognition (e.g., accents, dysarthria); even this variability could be prospectively addressed by training the speech recognition engine prior to testing. Timed prompts (e.g., “anything else?”) following a specified delay could be integrated to ensure patients have sufficient opportunity to respond. The limiting factor is the reliability and accuracy of speech recognition engines, which although they have made significant progress, are still prone to error; however, given that speech recognition algorithms have evolved to the point of being able to detect affective states (Conner-Simon, Gordon, & CSAIL, 2017) accuracy errors a likely to decline as the technology continues to evolve.

Assessing nonverbal learning and memory is perhaps even more challenging, because any automated scoring would need to utilize fuzzy algorithms to avoid unnecessarily rigid and stringent scoring; this too could potentially be customized based on patient population or settings (e.g., the degree of precision and accuracy required to satisfy scoring criteria could be reduced for movement disorder populations). One advantage of computer administered memory measures is the ability to develop multimodal paradigms that integrate simultaneous presentation of both verbal and nonverbal stimuli, which may be useful to disentangle memory for verbal information from memory for nonverbal information within the same task. To maximize utility and benefit, these new digitally-assisted assessment methods should be computer driven and developed with explicit utilization of technology. This does not have to be a screen per se, but the means by which the performance and behavioral data are captured should rely heavily on automated methods and minimize demands for administrator input once a task is initiated.

**Benefits: Beyond Scoring**

An obvious benefit of digitally-assisted assessments is the built-in standardization afforded via computerized administration. Presenting instructional sets and demonstrations via a digital interface, either through written or auditory text, or virtual-technician, computer generated instructions ensures compliance with standardization. Computerized instructions also can provide an opportunity to repeat sample items until task demands are fully understood. This increase in standardization by default, increases reproducibility and therefore reliability, and thus reduces the potential for administrator error. Automated scoring drastically reduces potential for scoring and data entry errors, and would facilitate real-time evaluation of standardized performance, both increasing efficiency. And while scoring programs already exist and are a staple in many clinics, they still require manual data entry after test completion, which is error prone, time consuming, and inefficient. These automations will cumulatively result in a substantial increase in overall clinical efficiency, which in the modern healthcare environment, is critical. Patient access to neuropsychological services will also improve since not only will results be available substantially faster (in some cases, immediately) than what is possible via analog practices, digitally-assisted assessments may allow for evaluating patients concurrently in some settings (e.g., inpatient service).

Another major advantage afforded by computerized testing is the ability to incorporate principles of modern psychometric theory including Item Response Theory (IRT), which has a number of advantages over classical test theory (Thomas, 2011). Computer adaptive testing (CAT) for example, leverages these IRT principles and is predicated on the ability to evaluate item-level performance in real time (Gibbons, Weiss, Frank, & Kupfer, 2016; Reise & Waller, 2009). This level of analysis requires the computational speed and power of a computer in order to be effective. Although some analog measures approximate an adaptive testing format (e.g., reversal rules), the starting point is so low in the difficulty curve that in most instances, a number of unnecessary items are administered that yield much less informative data. Instead, administration could start at the item associated with an average score for a given patient and then apply the reverse rule to administer the next easier item if the starting item is missed, or proceed to the next more difficult item if starting item is correct. This iterative process of adjusting the level of difficulty up or down based on performance on the previous item is the basic premise of adaptive testing, and ultimately leads to the concise item pool (and ability level) where patients obtain mostly correct scores. In addition to the increased precision, CAT paradigms have the major advantage of requiring significantly fewer items to determine the underlying ability level (Gibbons et al., 2008), improving efficiency even further. Although some domains might be more difficult for
IRT applications (e.g., memory, attention, working memory), a little creativity may foster development of novel IRT-based methods in these domains as well.

In order for adaptive measures to function, there needs to be a sufficiently large item bank to draw from. In addition to facilitating CAT applications, such large pools of items would have a number of other benefits. First and foremost, with sufficiently large number of potential items, several alternate forms could be developed. This would not only benefit longitudinal monitoring and re-evaluations, increasing the availability and flexibility of alternate forms would be another method by which test security could be maintained. Similarly, new “experimental” items could be regularly integrated and studied, which would allow for a continual rotation of new items and removal of outdated or “stale” items. IRT methods also allow for objective calibration and equating of test items (Thomas, 2011), which would improve development of tests for use in multiple languages and with diverse patient populations. These IRT approaches are common among standardized tests (e.g., Scholastic Aptitude Test, Examination for Professional Practice in Psychology) and would reduce (or potentially eliminate) the need for large-scale test revisions (i.e., test versions).

Points of Consideration

There are several notable issues that need be addressed, some of which are broadly applicable to technology integration, and some that are specific to in-laboratory assessments. The rigidity of a computer, while beneficial for standardization, can also be overly restrictive and impose constraints on testing that would not be present with analog measures. For example, testing of limits or adapting tests to salvage meaningful information when there are clear limitations impeding administration of standardized protocols would be far more difficult with a fully automated assessment. The underlying programs would have to be sufficiently adaptive to not only recognize a failure to maintain instructional set during testing, but to also provide appropriate reorienting cues where applicable. Automated computerized testing can also make it more difficult to revise an answer once it has been recorded. And while review of completed items might be acceptable in some circumstances, allowing test-takers to review an entire set of questions and answers is both inappropriate and inefficient, and impractical with CAT. For patients presenting with behavioral impulsivity, this may prove to be particularly problematic and would require careful consideration during task development.

Digital security is also a critical component of technological-based assessments that will absolutely need to be addressed prospectively and reviewed on an ongoing basis. In addition to the importance of maintaining the security and confidentiality of patient data, and adhering to state and federal regulations (e.g., HIPPA), obtaining informed consent from test takers to store data electronically may also be required. Although some individuals may understandably be uncomfortable with remote or cloud-based storage, 63% of respondents to a large international survey reported that they are comfortable with storing their medical records in The Cloud (Cisco, 2013). As cloud-based data storage of healthcare data becomes even more commonplace and generations raised with digital technology become the predominant users of healthcare services, the proportion of people comfortable with this storage format is likely to grow. Regardless of comfort level, any development efforts still need to include detailed plans for a response to a data breach or attempted breach. Although many institutions likely have a plan in place, individual practitioners and smaller clinics may not. There may also be international laws that need to be addressed depending on how tests are used and how data are stored, but ultimately, it is the clinicians responsibility to ensure that protected health information is, in fact, protected (Bauer et al., 2012).

The volume of data generated by these new tools will also likely be exponentially greater than that which is currently generated via analog measures. Ensuring that these data are stored in a safe and secure system that in addition to federally mandated privacy standards, is also compliant with Title 21 of the Code of Federal Regulations, Part 11 regarding electronic records will be critical. These regulations apply to technology systems that fall under the jurisdiction of the FDA and outline minimum standards to ensure that electronic records are maintained in a manner that is trustworthy, reliable, and equivalent to paper records (Electronic Records, 2016). Given that many cognitive assessment devices are now be subject to FDA regulations (Computerized Cognitive Assessment Aid, 2015), these Part 11 standards are likely applicable.

While not necessarily novel, integrating technology into existing tests and extracting new data points reflects the closest approximation to current practice standards, and thus minimizes the extent of change from current analog methods. Beginning with modernization of existing measures may thus facilitate transition towards a purely digital lab. Integrating technology into these analog measures will certainly augment assessment practices and enrich the available data, but pursuing this line of development will necessarily be limited by the confines of the tests because they were originally conceived in an analog framework (and many before the computer was even invented!) and fail to capitalize on some of the most significant benefits (e.g., CAT). Ideally, new data collection methods will be developed with technology as an integral and primary component as opposed to a secondary consideration. Without such constraints, the possibilities are infinitely greater (Table 2).
Remote and Portable Testing

One of the notable limitations of current analog measures is that they require individuals to travel to a laboratory or clinic to conduct an assessment. By integrating technology more explicitly and using tablets, phones, or highly portable computers, remote testing (e.g., internet based assessments) becomes a realistic possibility. Incorporating these measures with other technologies (e.g., structural and functional neuroimaging, electroencephalogram, transcranial direct-current stimulation) broadens applicability even further. Ideally, these methods will closely parallel the novel methods developed for in-lab assessments to facilitate portability across settings and comparability of results (e.g., following a brain injury patient from inpatient rehab to outpatient), and as such, many of the same benefits will be realized (e.g., standardization, efficiency, automation, etc.). The biggest distinction from the technology driven laboratory assessment is that remote assessment methods will by definition, require a purely computer or device-based interface. The options for user input and responding may also be somewhat limited to the standardly available devices (e.g., keyboard, mouse, audio, video, trackpads, pressure pads, joysticks, game controllers, smart pencils, and touch screens) because specialized input devices may not be readily available outside of the clinic. But as technology continues to evolve, more are likely to be developed.

One obvious advantage of internet or remote-based assessments is accessibility. Telemedicine is already an established means of healthcare and is becoming a more common means of clinical service provision for allied health providers (Klaassen, van Beijnum, & Hermens, 2016) and psychology (Luxtorn, Pruitt, & Osenbach, 2014). In addition to increasing access in previously underserved areas (Adjorlolo, 2015; Cullum, Weiner, Gehrmann, & Hynan, 2006; Grosch, Gottlieb, & Cullum, 2011), there is growing interest in utilizing telehealth even in areas with established provider networks, but targeting patients who have difficulty getting to a clinic (e.g., limited mobility, unreliable transportation). If remote assessment services are developed, the catchment area of each provider will expand considerably and the distribution of “underserved areas” will be drastically reduced, if not eliminated altogether. It may even be possible to one day run an entire practice via remote services, eliminating the need for a physical office space, which would reduce operating costs, savings that could be passed to patients.

To conduct a neuropsychological evaluation via telemedicine using our current tools would be challenging in most circumstances. Although interviewing would be straightforward, the majority of the current analog measures would require a remote technician for data collection, which is both expensive and impractical in most settings. Some analog measures would be amenable to remote assessment (i.e., those which rely primarily on auditory verbal stimuli and responses), and several studies have indeed shown an acceptable rate of agreement between remotely administered assessments and in-person assessments (e.g., Barton, Morris, Rothlind, & Yaffe, 2011; Cullum et al., 2006; Cullum, Hynan, Grosch, Parikh, & Weiner, 2014; Grosch, Weiner, Hynan, Shore, & Cullum, 2015; Wadsworth et al., 2016), though analog tasks with performance-based input demands would be more challenging.

While there are studies demonstrating the feasibility of conducting an assessment using videoconferencing tools, more work is necessary to demonstrate utility on a larger scale. More importantly, development efforts should focus on moving beyond a simple port of analog measures to a telemedicine format and instead on development of computer administered measures that can be implemented online via the internet. There is already evidence supporting the use of web-based data collection on a large scale (e.g., Germine et al., 2012), including the Automated Neuropsychological Assessment Metrics (ANAM; Kane & Reeves, 1997), which has been the subject of a multitude of empirical studies. Most of the commercially available brain training platforms are also entirely internet based, some with massive international user-bases. These training paradigms could be readily adapted to integrate an assessment component (which is already an implicit component of some platforms) through proper standardization and development of appropriate normative data. Many functional MRI tasks are also predicated on use of a remote computer-based assessment paradigm, and these tasks could similarly be standardized to establish their utility as a clinical measure.

Table 2. Novel technology based data collection tools and potential data points

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart pencil</td>
<td>Track the speed and consistency of written output</td>
</tr>
<tr>
<td>Proximity sensors/accelerometers</td>
<td>Placed into manipulated stimuli to record placement and errors</td>
</tr>
<tr>
<td>Pressure pads</td>
<td>Finger tapping, finger coordination</td>
</tr>
<tr>
<td>Step counter</td>
<td>Monitor number of steps for emerging parkinsonism</td>
</tr>
<tr>
<td>Smart watch</td>
<td>Measure galvanic skin response during test administration</td>
</tr>
<tr>
<td>Software</td>
<td></td>
</tr>
<tr>
<td>Speech recognition</td>
<td>Record grammatical errors and speech patterns</td>
</tr>
<tr>
<td>Typing tracker</td>
<td>Monitor keyboard dynamics and typing speed</td>
</tr>
<tr>
<td>Facial recognition</td>
<td>Evaluating affect during evaluation</td>
</tr>
</tbody>
</table>

Table 2 continues...
Points of Consideration

Many of the same concerns and challenges facing digitally-assisted laboratory assessments also apply to remote data collection efforts; however, there are several additional considerations that are unique to remote testing. One of the major issues to address is how to verify the identity of test takers. Identity verification is an ongoing issue throughout the digital world (particularly in financial sectors), and the failure of the password system (O’Brien, 2014; Swaby, 2012) has spawned entire industries dedicated to these and related issues. Biometric identification is one of the more promising emerging technologies and has recently found its way into mainstream hardware, including most smartphones (e.g., thumbprint readers), which demonstrates not only the feasibility, but also the practicality of biometric security. Retinal scans, facial recognition, keystroke dynamics, and voice recognition either independently or in combination may also hold promise. Combining biometric data with user-specific information (e.g., a unique one-time access code provided to the patient by the practitioner) increases the level of security even further. Two-level authentication is an additional means by which test security and usage could be controlled, because the access codes could require a prescription.

Validity. Another issue that must be addressed is how to assess the validity of obtained data. Beyond concerns of malingering or suboptimal performance, and all the standard psychometric properties that would of course need to be established prior to clinical use (e.g., construct validity, reliability), one unique aspect of remote assessments that needs to be addressed is random responding. In a clinic or laboratory setting, direct observations are an inherent component and this may be feasible via remote video monitoring in some settings because many devices include a video camera. For an examiner or technician to watch an entire evaluation though is highly impractical, and reviewing a delayed recording may be too late to intervene and redirect the examinee. Methods for real-time monitoring of engagement with testing need to be developed. Depending on the type of task used, random responding would be relatively easily detected, as item-level response patterns (e.g., consistency) could be monitored in real-time. In a CAT paradigm, random responding would be readily apparent because of an unpredictable pattern of errors that don’t follow a typical difficulty curve. When a pattern of random responding is detected, the program could respond in several ways including an alert to the examiner or re-introduction of task demands, or the task could be restarted or discontinued, and the video feed could be time-stamped for later review to qualitatively evaluate the behavior. Similar approaches may also be useful for detecting intentional underperformance. In conjunction with standard approaches to performance and symptom validity assessment, variability in responding along the difficulty curve may further support data invalidity due to underperformance.

Test security. Maintaining test security is a far more concerning and a much greater challenge. If an assessment is done entirely via online or remote methods, there is little protection of test items and content. There would be nothing in place to prevent someone from recording the assessment via external device or simply writing items down for future reference. Third-party observers now have the added risk of becoming third-party assistants, and even video feeds of the assessment would not be able to entirely safeguard against this. Ensuring item security could potentially be addressed through careful redesign of tests. For some measures, CAT paradigms could retain sufficiently large item banks that are drawn from at random, thus creating a new version each time. Tests could also be developed such that items are less discrete and less obvious as to where meaningful data are captured (e.g., rolling assessment point during continual stimulus presentation). For example, a continuous performance task where the speed of stimuli is gradually increased until a certain error threshold is reached, which is then used as the measure of ability. The task demands would be apparent, but there would not be any clearly distinct item content that could be recorded for later reference. While obviously these methods would not be feasible for the entire repository of available tests, they will work for some tasks. Protecting against an assistant, third-party observer, or external recording is much more difficult, as it would entail environmental manipulations from afar. While the prevalence of cheating is likely context-specific and some settings simply may not be amenable to remote assessment (e.g., forensic evaluations), until more effective measures to maintain test and item security are developed, fully remote assessments will be quite vulnerable.

Continual or High-frequency Data-capture

Since the Apple iPhone was released in 2007, there has been an exponential increase in mobile health research. The third largest indication was mental health and though much of this research targeted promotion of health behaviors, a sizeable portion focused on remote monitoring and data collection (Fiorelli, Diviani, & Schulz, 2013). Neuropsychology needs to integrate similar methods for capturing behavioral data via portable and wearable technology (e.g., tablets, phones, watches). Data reflecting cognitive functioning are readily available from everyday behaviors (i.e., “natural” data) and automated
approaches to collecting predefined data points from typical, everyday behaviors need to be defined and developed (i.e., “capture” it). We view this as conceptually distinct from classical “assessment” paradigms, which are explicit, intentional, and active methods by which behavioral data are elicited through narrowly defined methods. Ecological Momentary Assessments (EMA; Moskowitz & Young, 2006) for example, is an assessment method that utilizes remote collection methods to increase the frequency of data collection and Cognition Kit (CK) is a commercially available assessment platform that enables remote data collection built around wearable technology and is designed to “understand cognition in everyday life” (www.cognitionkit.com). Both EMA methods and the CK platform demonstrate the feasibility of remote assessments using currently available technology, and have notable advantages over traditional analog assessments (e.g., linking cognition and/or behavior to identifiable events). Though they do not fully address the primary limitations imposed by traditional measures because they still require an active engagement with the assessment device, which by definition alters the dynamic of the observation, they do offer a number of advantages.

Data capture methods to understand behavior are used extensively in other industries (e.g., market research, search engine optimization). Targeted direct marketing methods (e.g., pop up adds, banners, side bar ads) for example, are all driven by information gathered via PDC. The amount of data amassed on us and our behaviors via PDC methods each and every day is staggering, and nearly all of this is collected without our knowledge in an effort to preserve the “natural” interaction. PDC methods are entirely dependent on technology and simply cannot be implemented otherwise. Similar methods of recording/capturing the ways in which people interact with technology could be developed to understand more about real-world behavior. As a very simple example, the number of typing errors made and which side of the keyboard they occur on, typing speed, reading speed, speech errors and speech patterns (e.g., dysarthria) could all provide meaningful data. All of this can be easily quantified, recorded, and stored (if it isn’t already). And because it would be associated with demographic information of the device user (highlighting the importance of identity verification), it could be standardized against demographic information. Neuropsychology needs to leverage these methods and integrate them into clinical practice to understanding brain and behavior relationships via real-world observations.

Passive and continuous data collection methods have a tremendous number of advantages. First and foremost, the amount of data gathered on any given individual will be massive. Although along with the rich characterization of behavior, there will be a lot of “noise” to sift through in order to parse out the signal of interest, this could potentially be automated given sufficient forethought as to how the data will be used. Within these data, there will be an established baseline of an individual’s behavior. Such robust baseline data is incredibly useful because in addition to traditional normative reference to age and demographic peering, it allows for intra-individual comparisons over time, which in some instances is much more informative (e.g., neurodegenerative disease). When patterns of behavior emerge that reflect significant departures from historical baselines, they could automatically be flagged and/or trigger active collection of additional data points to enrich the dataset (e.g., contextual factors). Similarly, with identification of predefined outcomes (e.g., stroke, dementia), predictive algorithms could be developed via machine learning paradigms that would alert people to clinically meaningful changes in cognition or behaviors much earlier (i.e., prospective health monitoring). For example, changes in keystroke dynamics (e.g., typing speed, how long keys are depressed) or the frequency of spelling errors that deviate substantially from historical trends or normal age-related changes might be indicative of a stroke. For those who dictate messages or emails, a change in the frequency of pauses could be used to detect word finding errors. An increase in grammatical errors may signify an emerging progressive aphasia. Conversely, failure to keep pace with normal age-related developments among adolescents might signal developmental delay or trigger the need to update rehabilitation efforts following acute injury. It may even be possible to detect emerging depressive symptoms or suicidality based on the patterns of Tweets or characteristics of Instagram posts, a project which is already in development (see Reece & Danforth, 2016).

Another advantage of passive or continual data collection is that behavioral data are captured in real-time. Current analog methods are restricted to an isolated time point, providing a static estimate of a dynamic ability that varies as a function of environmental factors and individual context. Monitoring cognition in real time allows for continuous assessment to establish a more robust estimate of ability, and if integrated with active data collection methods to identify additional information about the local environment, it would be possible to see how cognition varies as a function of context (e.g., reading speed in the library vs. the coffee shop). Even more important, longitudinal monitoring with more than annual precision will be possible, which is particularly useful for disentangling potential variance attributable to extraneous factors (e.g., patient came in for follow-up testing following a poor nights sleep). Longitudinal monitoring with such high frequency would be particularly relevant for clinical trials and monitoring treatment effects, as well as observing recovery trajectories following injury or rates of decline in neurodegenerative disease. The information from longitudinal data collection like this could be made available in a simple and digestible form to primary treating providers (e.g., data download during an annual wellness visit for older adults), which may prompt for more detailed active cognitive assessments or updating of treatment plans.
Ecological validity is in a related vein and has been defined as “functional and predictive relationship between the patient’s performance on a set of neuropsychological tests and the patient’s behavior in a variety of real world settings” (Sbordone, 1996; p. 15). Most neuropsychological tests were not developed for this purpose and not surprisingly, have demonstrated only a moderate level of validity for predicting functional “real-world” behaviors (Chaytor & Schmitter-Edgecombe, 2003). Instead, predictions about behavioral functioning outside the lab are made on the basis of assessing the cognitive abilities thought to underlie these behaviors (Marcotte, Scott, Kamat, & Heaton, 2010). If methods were developed that allow for quantifying behavior in the real world, in real time, alongside information about the local context, ecological validity of cognitive assessments may improve substantially. Having such data available may also inform better treatment recommendations and thus improve patient outcomes, because the evidence upon which recommendations are based upon would stem directly from real-world, real-time observations. If integrated with remote monitoring (e.g., automated alerts to treating providers), treatment recommendations could be updated much sooner than otherwise feasible using traditional assessment practices, which would also improve patient outcomes.

Points of Consideration

Integrating PDC methods via portable and wearable technology is no small feat, and there are a multitude of issues to address prospectively. While many individuals are the sole user of a given device, some devices may be shared or borrowed. Confirming the identity of the user will be challenging, even more so than with remote assessments, because of the inherent portability of these devices. It may be possible to temporarily turn off data collection and behavior monitoring when a device is used by another individual, though this would have to be automated to some extent. Otherwise, the target user may inadvertently forget to resume behavior monitoring. In addition to establishing the basic psychometric properties of PDC methods, linking the observed behaviors to brain functioning will be an important part of development. Relatedly, identifying how these data map onto existing cognitive constructs established by analog measures will also be critical.

There are also issues of informed consent. It would be highly unethical to install a monitoring application or software onto an individual’s device without their consent. One possibility is that the data are collected automatically and consent of the patient is required to download the data for clinical use. Otherwise, the data are simply stored until needed. If concerns of aphasia were raised (e.g., PPA) or a patient had a stroke for example, the historical data could be accessed by their providers after providing consent, but not before. Alternatively, any data collection efforts could require an individual to “opt in” instead of “opting out,” which would thus inherently obtain informed consent and since the passive response is non-participation. A tiered participation option may also be possible, allowing individuals to specify the extent of information gathered and with whom it may be shared. One of the concerns with a priori consent is that for some individuals, the knowledge of being “observed” may change the dynamic between the person and the technology and thus influence the data; however, the true influence of this knowledge on behavior reflects an empirical question of interest that should be addressed during development of these methods. Another possibility is that installation of such background programs could only be done at the direction of a patient’s healthcare provider, under the care of an appropriately trained clinical neuropsychologist, comparable to current prescription-based referrals. This approach may provide an additional safeguard against unscrupulous use of such technology, or use by unqualified providers. Regardless, data access would have to be heavily restricted to protect both patient privacy and test security and the same data security issues previously discussed would also apply.

Barriers to Technology Integration

Perhaps the greatest challenge in adopting more technologically-driven measures of cognitive functioning is actually getting clinicians to transition away from existing tried-and-true analog measures. One of the first and most critical things to understand is what have historically been the barriers to adaptation of currently available technologically advanced methods. This may require some quality market research among the field, done in an effort to understand the current concerns neuropsychologists have about digital testing. A major issue that has plagued the current computerized assessments is concern about both the quantity and quality of independent validation data (Bauer et al., 2012), and establishing the same minimum psychometric standards will certainly remain essential. One advantage afforded to novel technology driven methods is the wealth of data and experience base available from analog measures, which will serve as one source of criterion validity. Since FDA requirements now necessitate that novel computerized neurocognitive assessment devices demonstrate sufficient psychometric validity and reliability and an appropriate normative database prior to approval for use (Computerized Cognitive Assessment Aid, 2015), the psychometric floor has already been set. In looking beyond neuropsychology-specific issues and thinking about general considerations that may have been raised in other fields, there are several additional potential barriers that come
to mind. At its core, potential users need to be convinced of the benefits of technology-driven data collection methods and assessments above and beyond current analog measures. To what extent this utility needs to be demonstrated and what it will take to convince users to switch, however, remains to be seen.

Another concern often informally raised among the neuropsychology community is that integrating technology will render the clinician obsolete and ultimately lead to the demise or dilution of the field as a subspecialty as it gets “taken over” by computer-based tasks. The fear is that technology increases the feasibility of fully automated assessments and computer-generated reports (i.e., a clinician in a box), which would eliminate the need for human interpretation of the data and thus allow for assessment by providers without appropriate training and expertise. Misuse of automated testing is of course a possibility, especially if the field relinquishes control over assessment methods and tasks others outside of neuropsychology with the development and evolution of data collection methods. Through such passivity, the field will lose its position and voice and the neuropsychologist will be relegated to the role of a technician; there are emerging indicators that this is already happening. However, if we retain ownership of cognitive and behavioral assessment and drive the development and evolution of novel methods to capture behavioral data, we will not lose our seat at the table and retain our position as experts in measurement of human behavior.

Conversely, if properly developed, the expertise of the well-trained neuropsychologist will become that much more valuable. For example, if cognitive and behavioral data become so prolific and readily available, there may actually be greater demand for skilled interpretation. There are growing concerns among radiologists that with the advent of automated scanning protocols, they will be replaced by computer software (Chockley & Emanuel, 2016). However, even with the advent of automated volumetric analysis and artificial intelligence, there are still demands for human interpretation. The more likely scenario is that radiologists simply need to evolve their practice (Jha & Topol, 2016), which is what the field of neuropsychology also needs to do. For a proper neuropsychological evaluation, there will still be a demand for a proper clinical interview, test selection, record review, and with the proliferation of passive data, integration of all this information into a cohesive patient profile. It is important to remember that integrating technology merely advances the ways in cognitive and behavioral data are collected. It does not by default, mitigate the need for a skilled clinician to guide the evaluation process and interpret the obtained data.

There are also notable switching costs associated with transitioning to a digital clinic or lab. In addition to the financial costs associated with buying the hardware and software for a new test, there are also educational costs associated with learning a new suite of tests. Unfortunately, the financial costs are real and unavoidable. However, the educational switching costs may be less significant than one might think, depending on how new tools are developed. If integrated from the earliest training opportunity (i.e., graduate school), these would be a core component of common knowledge. Future trainees would be immersed in a technologically driven practice from the start, because at present, one barrier to integrating technology is limited exposure at the training level (Rabin et al., 2016). And if developed using existing tests as the gold standard comparison and basis for validation, the knowledge gap between digital and analog measures would be significantly lessened. At the end of that day, there will of course be a learning curve. However, this is universally true with any new measure and under no circumstances should new learning be a barrier to the growth and development of the field.

Some clinicians may feel that integration of technology will result in a loss of qualitative behavioral data (e.g., Boston Process Approach). This qualitative data is often used to contextualize the quantitative data and facilitate interpretation in light of factors “in the room” that may not be readily apparent in the data. With some of the proposed methods there will be no opportunity for direct observation of behavior, and some may worry that these data will be lost if assessment methods are automated. For assessments done in the lab or clinic, there will still be opportunity for direct, in vivo observations, and integration of technology would allow for preservation of these data for future reference. Similarly, most computers and phones have a built-in camera that would also allow for behavioral observations during remote assessments. For passive data capture methods, technology would facilitate collection of additional data points that could systematically capture and quantify the underlying process.

Another frequently raised concern about integrating technology is that there is a substantial portion of test users and test-takers who are uncomfortable or intimidated by technology (i.e., it is not uncommon to hear from patients: “I’m not good with computers”). This is certainly an issue for some individuals, especially older cohorts, though this does not likely represent the majority of test-takers. Furthermore, it could be easily argued that this is essentially an irrelevant argument now with a fast-approaching expiration date. Within the next decade, such concerns will likely be completely unfounded. For the few for whom this is an issue, proper development of administration procedures (e.g., video tutorials, speech recognition), should nullify any such concerns. What is far more concerning is that the inverse will eventually be true (i.e., that not using computers or tablets will be unusual) and that current analog methods will be strikingly foreign. Many schools today rely on tablets and computers in the classroom as a standard part of their education, and some curricula now include coding and programming. As the younger and future generations age, testing via analog measures without a computer may be experienced similarly to those for whom technology is intimidating.
Remaining Unknowns

There are a tremendous number of unanswered questions and remaining uncertainties, and the thoughts presented here are by no means intended to be comprehensive. Many of the practicalities raised in recent position statements (e.g., psychometric validation, review and approval by the Food and Drug Administration; Bauer et al., 2012) have not been discussed in detail, and there are additional settings and opportunities for integration of technology that have not been addressed (e.g., technology based interventions). By now, however, it should be readily apparent that there are a tremendous number of advantages to technologically driven data collection methods. It is our belief that the comprehensive assessments of the future will likely entail a combination of lab-based assessments, remote assessments, and passive data capture. Though there remain a number of very real issues that need to be considered, none are so significant as to preclude evolution of the field. And despite the number of uncertainties ahead, the probability that assessment is going to change one way or another is clear. The real question is no longer “if” technology will become a mainstream staple of cognitive measurement, it is “when” its going to happen and whether or not these changes are championed from within the field or again relegated to third-parties. While we must approach the integration of technology from a multidisciplinary mentality, incorporating the fields of bioinformatics, computer science, engineering, gaming, and business with our expertise in cognition, psychometrics, and neuroscience, it is critical that we overcome the current inertia and assert our position as experts in quantification of human behavior lest we get pushed to the wayside. The development of technology driven tools for collection of behavioral data should become a general priority of the field that is guided in part by current clinical demands, as well as, anticipated demands that may emerge in the future. The specific priorities, however, need to be established via consensus process that not only involves a diverse representation of the field, but other disciplines as well (e.g., computer science and informatics). If done properly, the most significant change to the field will be evolution of the tools neuropsychologists use to collect and capture data, though the core components of clinical practice should remain comparable over time. Our expertise will unequivocally be required to drive continued development of novel data collection methods and interpret the data in the context of human brain and behavior relationships. The overarching goal is to improve patient outcomes and make evidence-based contributions to patient care, advance the scientific knowledgebase of brain and behavior relationships, and facilitate ongoing development of the field.

Conflict of Interest

None declared.

References


