The development of an online neuropsychological test battery

The Amsterdam Cognition Scan

Feenstra, H.E.M.

Chapter 2

Online Cognition: Factors Facilitating Reliable Online Neuropsychological Test Results

Helleen E. M. Feenstra
Ivar E. Vermeulen
Jaap M. J. Murre
Sanne B. Schagen

The Clinical Neuropsychologist, 2017; 31(1), 59–84.

ABSTRACT

Objective

Online neuropsychological test batteries could allow for large-scale cognitive data collection in clinical studies. However, the few online neuropsychological test batteries that are currently available often still require supervision or lack proper psychometric evaluation. In this paper, we have outlined prerequisites for proper development and use of online neuropsychological tests, with the focus on reliable measurement of cognitive function in an unmonitored setting.

Method

First, we identified several technical, contextual, and psychological factors that should be taken into account in order to facilitate reliable test results of online tests in the unmonitored setting. Second, we outlined a methodology of quality assurance needed in order to obtain reliable cognitive data in the long run.

Results

Based on factors that distinguish the online unmonitored test setting from the traditional face-to-face setting, we provide a set of basic requirements and suggestions for optimal development and use of unmonitored online neuropsychological tests, including suggestions on acquiring reliability, validity, and norm scores.

Conclusions

When properly addressing factors that could hamper reliable test results during development and use, online neuropsychological tests could aid large-scale data collection for clinical studies in the future. Investment in both proper development of online neuropsychological test platforms and the performance of accompanying psychometric studies is currently required.

INTRODUCTION

Neuropsychological assessments are the first-choice method to obtain objective measurements of cognitive functioning (Gates & Kochan, 2015; Lezak et al., 2004), but their time-consuming and labor-intensive nature limit large-scale data collection for clinical studies (Caine et al., 2012; van de Weijer-Bergsma, Kroesbergen, Prast, & Van Luit, 2015). Many current research topics, such as age-related cognitive decline, eHealth interventions, and genetic-cognitive associations, require accessible and efficient tests for cognitive measurements in large samples (Aalbers, Baars, Olde Rikkert, & Kessels, 2013; Haworth et al., 2007; Wild, Howieson, Webbe, Seelye, & Kaye, 2008). This is especially the case since clinical populations are often dispersed by highly varied individual, disease, and treatment characteristics that might moderate cognitive functioning. For example, the risk factors for cognitive decline following cancer treatment can, in addition to the nature of received treatments, depend on various patient characteristics such as age, disease subtype, disease stage, cognitive reserve, genetic risk factors, and comorbid conditions (Wefel et al., 2011). New developments such as the advent of "personalized medicine", in which treatment is more tailored toward individual disease characteristics, further increase variability of moderating factors. To be able to assess the differential effects of this growing number of moderating variables on cognitive functioning in clinical populations with adequate statistical power, increasingly large sample sizes are needed. This is no longer feasible using supervised face-to-face neuropsychological testing alone. To keep research on cognitive functioning up to speed with new developments in treatment and diagnostics, tools that enable collecting reliable data in an unmonitored setting are becoming increasingly important (Aalbers et al., 2013; Gates & Kochan, 2015; Gigler, Blomeke, Shatil, Weintraub, & Reber, 2013; Haworth et al., 2007). With the rise of computerized testing as a means to assess cognitive functioning in clinical practice (Bauer et al., 2012; Iverson, Brooks, Ashton, Johnson, & Gualtieri, 2009), and the current widespread availability of computers and Internet connections in patients' households (http://www.internetlivestats.com/internet-users/#byregion; http://www.internetworldstats.com/stats.htm), an efficient and economical way to significantly increase patient sample sizes for studies on cognitive functioning would be to employ tools for *online* neuropsychological testing. Since the use of online neuropsychological tests is still in its infancy, there is an ongoing search to find its optimal applications while dealing with several unresolved issues. In this paper, we will discuss the advantages and challenges of online testing to study cognition in clinical studies and at the same time provide guiding information for the development and use of online neuropsychological tests.

Advantages of online testing

there are several possible research settings to gather information from patients using neuropsychological tests: (1) a test leader conducts paper-and-pencil tests face-to-face; (2) a test leader conducts paper-and-pencil tests from a distance via telephone or video call; (3) patients take a computerized test under supervision; (4) patients take a computerized test independently, usually after face-to-face instructions; (5) patients take an online computerized test (on site or at home) under supervision; and (6) patients take an online computerized test (on site or at home) unsupervised. A neuropsychological assessment may consist of combinations of settings, such as

a face-to-face assessment with several paper-and-pencil tests and one computerized test. Online testing (setting 5 and, more commonly, 6) is a subtype of computerized testing (settings 3 and 4), the latter having certain advantages over traditional paper-and-pencil testing (settings 1 and 2). First, through computerized testing, test scores are acquired more objectively and precisely. Stimulus presentation, scoring, data transfer, and data storage are standardized, which lowers the risk of human errors and increases measurement precision (Barak & English, 2002; Bilder, 2011; Parsey & Schmitter-Edgecombe, 2013; Reips, 2002c). Second, computerized tests can benefit from additional technical possibilities such as the presentation of engaging multimedia stimuli, the measurement of additional observations (e.g. reaction times), adaptive testing, auto-archiving, and automated normative data (Barak & English, 2002; Bauer et al., 2012; Naglieri et al., 2004; Schatz & Browndyke, 2002). Third, paper and money may be saved in testing materials (e.g. tests and scoring forms) and by having patients complete tests simultaneously (Naglieri et al., 2004; Reips, 2002d).

Additionally, online testing has specific advantages over computerized testing, as it poses few restrictions on timing and location of assessment and is centrally managed on a web server (Barak & Buchanan, 2003; Birnbaum, 2004; Caine et al., 2012; Naglieri et al., 2004; Reips, 2002d). Participants can access online test platforms 24 h a day, from home or other convenient locations. Therefore, patients from remote locations can also participate and patients with mobility problems or busy lifestyles do not have to travel to a specific test location (Birnbaum, 2004; Buchanan & Smith, 1999; Caine et al., 2012; Germine et al., 2012; Naglieri et al., 2004; Reips, 2002d). Furthermore, there are fewer costs for test locations (Naglieri et al., 2004; Reips, 2002d). The online access also enables researchers to access data wherever there is a secure Internet connection (Reips, 2002d). Because of the central management and administration of online testing platforms, data are collected in a single database, which greatly facilitates generating normative data (Barak & English, 2002). Moreover, software updates, bug fixes, and test adaptations are implemented centrally, which ensures that all users work with the latest version of the online testing platform (Barak & Buchanan, 2003).

Finally, unsupervised testing has particular advantages over testing under supervision of a test leader. First, labor, training, and travel costs for test leaders are saved (Bauer et al., 2012; Buchanan & Smith, 1999; Caine et al., 2012; Naglieri et al., 2004; Reips, 2002c; van de Weijer-Bergsma et al., 2015). Second, and importantly, reliability of neuropsychological tests may benefit from the absence of a test leader in terms of reduced experimenter effects. While each face-to-face assessment is influenced by personal behavior of the test leader (different feedback, instructions, observations, participant-test leader interaction), the computerized test has only limited personalized features (Reips, 2000). Also, people may be more willing to answer personal questions candidly in the absence of a test leader (Gunter, Nicholas, Huntington, & Williams, 2002; Reips, 2002b, 2002c, 2002d).

Advantages of both computerized testing and unsupervised online testing are displayed in Table 1. Because unsupervised online testing leverages the advantages of computerized testing and

adds to that the ease of large-scale online recruitment, data collection, and data and software management, as well as the economies and improved reliability of unsupervised testing, it may prove an interesting assessment tool for neuropsychological research. However, there are also challenges to online neuropsychological testing; some of these challenges are typical for online testing, and some are common for all neuropsychological tests. These challenges, and how to overcome them, will be discussed in the following sections.

Table 1. Advantages of computerized testing and unsupervised online testing.

Advantages of computerized testing (references)
• Standardization: presentation of stimuli and feedback, data storage, and scoring (Barak & English, 2002; Bilder, 2011; Parsey & Schmitter-Edgecombe, 2013; Reips, 2002b)
• Extra data points (e.g. reactions times & mouse coordinates) to measure multiple dimensions of performance at levels unattainable for human observers (Bauer et al., 2012; Schatz & Browndyke, 2002)
• Reduced costs for test administration and scoring (Bauer et al., 2012; French, 1986)
• Reduced time for preparation of experimental setup, scoring, analysis of response, and data transfer (Barak & Buchanan, 2003; Barak & English, 2002; Bauer et al., 2012; Schatz & Browndyke, 2002)
• Possibility of multi-media presentation: capture interest of participant and test abilities which are difficult to test (Gunter et al., 2002; Naglieri et al., 2004; Schatz & Browndyke, 2002)
• Possibility of adaptive testing (more efficient use of time) (Bauer et al., 2012)
• Flexibility to administer tests in different languages (Bauer et al., 2012)
• Ability to integrate and automate interpretive algorithms (e.g. determining impairment, statistically reliable change) (Barak & English, 2002; Bauer et al., 2012)
• Automated data export for research purposes (Bauer et al., 2012)
• Mandatory fields prevent oversights or omissions (Gunter et al., 2002)
• Automatic gathering of normative data (Barak & English, 2002; Schatz & Browndyke, 2002)
• Flexibility to generate alternate forms for repeated testing (Gualtieri & Johnson, 2006)
• Participants can feel more comfortable revealing sensitive data about themselves to computers (Booth-Kewley, Larson, & Miyoshi, 2007)
• Opportunities in developing more ecologically valid tests (Smit et al., 2013)
• Reduce disposable materials (test material and storage) (Barak & English, 2002; Schatz & Browndyke, 2002)
• One system for all elements of the test battery and for the whole process from administration to interpretation (Russell, 2011)
Advantages of unsupervised online testing (references)
• Efficiency (time, materials, and money) (Birnbaum, 2004; Naglieri et al., 2004; Reips, 2002c)
• No experimenter effect (variability caused by test leader) (Birnbaum, 2004; Naglieri et al., 2004; Reips, 2002c)
• Low/ no costs for personnel, equipment, location, and travel (Bauer et al., 2012; Buchanan & Smith, 1999; Caine et al., 2012; Naglieri et al., 2004; Reips, 2002c)
• Easy access: possibility of long-distance testing (Caine et al., 2012; Naglieri et al., 2004; Reips, 2002c)
• Heterogeneous samples (culturally diverse and hard to reach participants) and generalizability (Birnbaum, 2004; Buchanan & Smith, 1999; Germine et al., 2012; Reips, 2002c; van Steenbergen & Bocanegra, 2015)
• Fewer political/geographical borders (Barak & English, 2002)
• No time constraints (Barak & Buchanan, 2003; Reips, 2002c)
• Simultaneous participation (Reips, 2002c)
• Easy updates and certainty that latest version is used (Barak & Buchanan, 2003)
• Richer responses to open-ended questions and less social desirability in answers (Gunter et al., 2002; Reips, 2002c)
• Convenience of home testing and possibly less stress (Darby et al., 2014)
• Norms can continuously and immediately be updated (Barak & English, 2002)
• No need to download or install software (Hansen et al., 2015)
• Closer approximation of real-life situation (van de Weijer-Bergsma et al., 2015)
• Greater external validity through greater technical variance (Reips, 2002c)

Challenges of online neuropsychological testing

Even though neuropsychological computerized tests have been available since the 1980s (Barak & English, 2002) and use of computerized tests for neuropsychological assessments has increased among researchers and clinicians in recent years (Parsey & Schmitter-Edgecombe, 2013), the use of online tests has not been picked up as much in neuropsychology as in other fields of psychology. While general psychology has embraced large-scale online testing in studies on, for example, memory of daily news (Meeter, Murre, & Janssen, 2005) and social attitudes (https://implicit.harvard.edu/implicit/), and numerous online questionnaires are available for clinical screening and monitoring (e.g. (Cella et al., 2010; Medalia, Lim, & Erlanger, 2005)), neuropsychological testing has undergone little fundamental change over the last century (Bilder, 2011; Schatz & Browndyke, 2002). According to Parsey and Schmitter-Edgecombe ((2013), p. 1328) 'despite supportive findings for […] technology-based assessments, there is resistance in the field of neuropsychology to adopt additional measures that incorporate technology components'. For objective information on cognitive functioning, paper-and-pencil cognitive tests are still the first-choice method among many researchers and practitioners.

Neuropsychological assessments aim to assess optimal cognitive performance using standardized, reliable, and validated measurement tools. A general concern about online neuropsychological assessments is that they bring about various technological and contextual constraints, which could hamper optimal test performance and therefore fail to represent "true performance". Neuropsychological tests, and tests based on reaction times especially, are sensitive to testing environment, which is difficult to control online (Crump, McDonnell, & Gureckis, 2013).

Despite the field's concerns with respect to online assessments, several online neuropsychological tests are available at the moment (Iverson et al., 2009). Table 2 provides an overview of established neuropsychological test batteries developed for online use. In addition, systems for development of your own online test (e.g. Inquisit 4 web) and several stand-alone online neuropsychological tests are available as well. However, when focusing on their psychometric qualities, the widespread use of many of these online cognitive tests is hampered by concerns about the data they elicit. First, reliability and validity of the tests are often not properly studied and reported (Gates & Kochan, 2015; Naglieri et al., 2004). Second, tests are usually not designed to be conducted in an unmonitored setting (i.e. without a test leader present) or on a large variety of hardware and software, or they lack an administrative user interface to enable large-scale assessment (Hansen, Haferstrom, Brunner, Lehn, & Haberg, 2015). And third, normative data are often not available or merely based on (non-equivalent) offline assessments (Bauer et al., 2012; Germine et al., 2012). Furthermore, whether or not online tests are suitable for specific target groups, such as people with poor computer skills or severe cognitive problems, is often not explicitly studied or mentioned.

Concerns about the applicability of online testing have resulted in several guidelines on the development and use of online tests, both for general psychology and for neuropsychology. Reips (2002c) has provided guidelines for general online testing, Schlegel and Gilliland (2007)

for computerized psychological test batteries, Thompson and Weiss (2011) for computerized adaptive testing, and Bauer and colleagues (2012) for computerized neuropsychological testing.

Table 2. Established online neuropsychological test batteries.

Test	Content	Norms: sample (n) age range	Results reported	Citation
Unsupervised tests				
BAM-COG	Working memory, executive function, episodic memory, visuospatial short-term memory (4 tests)	Healthy adults (397) 40–85	Convergent validity: r=.20 to .67, p<.15; alternate form reliability: ICC = .17 to .65	Aalbers et al., 2013
CFT	Episodic memory, executive functioning, processing speed (4 tests; cognitive screening)	Adults without memory complaints/ AD (195) 50–65	Concurrent validity computerized version: r=.39 to .74, p<.005; internal consistency computerized-online: α=.73	Trustram Eve & de Jager, 2014
Cogstate brief battery	Processing speed, attention, visual learning and working memory (4 tests; cognitive screening)	None	Concurrent validity computerized version: r=.49 to .83	Maruff et al., 2009
CST	Verbal fluency, visual-spatial/ executive functioning, working memory, attention, orientation, processing speed. (6 tests; aD screening)	None	Concurrent validity total score: r=.56, p<.001; Test-retest (6 weeks): Cronbach's α=.76, F(47)=4.17, p<.00 sensitivity: 99%; specificity: 95%	Dougherty et al., 2010
Memoro	Verbal memory, spatial memory, working memory, processing speed (4 tests)	None	Concurrent validity: r=.49 to .63, p<.01	Hansen et al., 2015
WebNeuro	Sensorimotor, memory, executive planning, attention, emotion perception (11 tests; broad cognitive functioning)	None, only from offline version	Concurrent validity (compared with IntegNeuro): r=.43 to .87	Silverstein et al., 2007
Supervised tests				
CNS vital signs	Verbal and visual memory, processing speed, working memory, motor speed, reaction time, executive functioning, sustained attention, reasoning, social cognition (10 normed tests)	Healthy volunteers (1,069) 7–90	Test-retest reliability (average 62 days; n=99): r=.31 to .87	Gualtieri & Johnson, 2006; http://www.cnsvs.com/ academicresearch. html
ImPACT	Verbal memory, visual memory, processing speed, reaction time (6 tests)	Adolescents (4,500) 13–21	Concurrent validity composite scores: r=.18 to .43, sensitivity: 81.9%; specificity: 89.4%	Allen & Gfeller, 2011; Henry & Sandel, 2015; Schatz, Pardini, Lovell, Collins, & Podell, 2006
Vienna test system – neuro	Approximately 120 psychological tests; few online available neuropsychological tests	None, only from offline version		http://www. schuhfried. com/ viennatestsystem10/ vienna-test-systemvts/

In addition, the American Psychological Association has formed a task force to study online psychological testing (Naglieri et al., 2004), which has pointed out several important issues for proper use. In this article, we will elaborate on these existing guidelines, focusing on online neuropsychological testing in an unmonitored setting, and aim to identify factors facilitating

optimal test performance in this setting. In this analysis, we will extend prior studies, which focused mostly on technical issues, data safety, and the role of computer literacy (Bauer et al., 2012; Gates & Kochan, 2015; Germine et al., 2012; Schlegel & Gilliland, 2007). First, we will explore technical, contextual, and psychological factors that distinguish the online unmonitored test setting from the traditional face-to-face setting and provide suggestions for optimal development and use of online tests. Second, we will address the need for and methodology of quality assurance (norms, psychometric properties, and support for users) of online neuropsychological tests. In doing so, we aim to provide up-to-date and evidence-based guidance for researchers interested in conducting and designing online unmonitored neuropsychological assessments. Well-founded information on the strengths and weaknesses of online cognitive research data collection can place current concerns into context and clarify when online testing is suitable, and when it is not.

TOWARD RELIABLE COGNITIVE DATA FROM AN UNMONITORED SETTING

The main characteristic that distinguishes online measurement tools from offline tools is that online tests will usually take place in an unmonitored setting. In contrast to face-to-face and computerized testing, participants of online tests typically lack instruction by, and supervision of, a test leader. According to the American Academy of Clinical Neuropsychology, the test leader's task is 'to establish a physically and interpersonally comfortable testing environment, with the goal of minimizing anxiety, resistance, physical discomfort, or other factors that may interfere with optimal motivation and effort' (AACN, 2007); p. 222). By attending to the patient's needs, abilities, and limitations, the test leader shapes the physical and psychological context to facilitate optimal test performance (Lezak et al., 2004). In addition, it is the task of the test leader to provide all test materials and make sure tests are properly understood before the actual assessment starts. Certainly, many factors may interfere with patients' motivation and performance during test situations. This can result in noisy data and, possibly even, dropouts. In the absence of a test leader, ways to handle these issues may differ from test situations with a test leader present. In the following paragraphs, we will review a number of these issues and suggest solutions on how to handle them in order to be able to collect reliable cognitive research data in an unmonitored setting. We will distinguish technical, contextual, and psychological issues.

Technical issues in online neuropsychological testing

Data collected through online testing may be affected by participants' computer configuration and Internet connection (Bauer et al., 2012; Reips, 2002d; Schlegel & Gilliland, 2007). Variations in software, Internet speed, processing power, and external hardware may influence (1) stimulus presentation, (2) participant behavior, (3) input processing, and (4) recorded performance. Despite the standardized characteristics of computerized testing, such variations could produce noise in the cognitive measurements. Technical variability thus can lead to unreliable data, which is poorly comparable with the tests' normative database. For example, in a test on reaction times, the computer's processing speed may result in an unknown delay of the appearance of a stimulus, influencing the registered time of the button press (participant behavior), processed with again an unknown delay (input processing), and, finally, an affected test score. Especially for

measurements based on reaction times, but also for example sound, video, or vector graphics, it is critical that each participant has access to a computer and Internet connection that meet the requirements to accurately represent stimuli and capture responses.

Below, we review a number of technical issues that researchers should be aware of. First and foremost, choices in the design and programming of an online neuropsychological test should minimize the effect of technical variation on test results. Subsequently, it is important to (1) thoroughly test the programmed tasks for functionality on all major systems (types of operating system and device) and browsers and (2) communicate technical minimum requirements (Bauer et al., 2012; Birnbaum, 2004). Fortunately, nowadays online cross-system and cross-browser testing services are available (e.g. Browserstack (https://www.browserstack.com/) and Sauce Labs (https://saucelabs.com/)) that allow remote testing of online applications, using various types and versions of browsers on real (i.e. not emulated) hardware. Testing can be performed manually (developer, researcher, or participant takes the online test) or automated (test taking is programmed in order to run the test in a standardized fashion and, possibly, simultaneously on several combinations of technical settings). A comprehensive test should be done at launch, followed by periodical testing, to ensure that applications run well on new systems and (versions of) browsers. Running the application on older, low-end systems and browsers will determine technical limits. Minimum requirements can consequently be communicated to researchers (specified in user manuals) and participants (during recruitment and in general test instructions) in order to control the technical setting. The participant instructions should be clear before the start of the test. Additionally, during the test, certain technical conditions can be checked on minimum requirements electronically (resulting in automated participant feedback where needed).

Server-side vs. client-side processing

Computer configurations affect data recording differently depending on the processing method (Bauer et al., 2012). Online tests run in web browsers and are based on a combination of server-side and client-side processing. Server-side processing fetches data from the database and creates the web page served in the browser; client-side (= browser) processing allows for dynamic web pages that respond directly to the user without consulting the server. The trend in recent years has been to shift more and more processing to the client side. This is enabled by better browsers, faster hardware, and more uniform standards for the computer languages and frameworks in which client-side processing is defined (mainly JavaScript, HTML, and CSS), limiting the influence by technical variance from participants. In particular, HTML5 stimulus presentation can approach millisecond accuracy (Garaizar, Vadillo, & Lopez-de-Ipiña, 2014). In case of server-side programming, stimulus presentation may be affected by the quality and speed of participants' Internet connection. In order to obviate this potential problem, browser "plug-ins"—additional software components such as Flash, Java, or proprietary plug-ins—may be used to run the script on the client side. This may have the disadvantage that not all plug-ins are supported by all browsers. However, recent developments in web standards make most of these plug-ins superfluous on the most recent generation of browsers.

Online tests are thus more and more based on client-side processing, making them somewhat more susceptible to participants' computer configuration (e.g. plug-in settings and processing speed), but, at the same time, less susceptible to Internet connection and server load. Server-side programming and client-side programming can be combined for optimal reliability and usability of a test. For example, client-side programming can be used for optimal stimulus timing and server-side programming for safe data transfer to the database. If programmed well, online tests running on modern browsers may closely mimic the behavior of locally installed computerized tests, except when they make extreme demands on the hardware (e.g. fast-moving, high-definition 3D graphics). Knowing that client-side programming is more influenced by computer configuration, predominantly used client-side elements can be tested on performance and minimum requirements by using the aforementioned tests for system and browser suitability.

Software standards

The use of different versions of test software can cause variability in recorded data (Plant & Quinlan, 2013). Variability may also arise from variations in participants' software (e.g. operating systems, plug-ins, and browsers). In order to make a test reliable and available to as many participants as possible, software configurations should fulfill certain criteria. This can be achieved, first, by testing browser versions across operating systems to rule out bugs and to ensure consistency of stimulus presentation and processing of input (ITC, 2006; Reips, 2002c). As some browsers have regular automatic updates that may influence the behavior of tests, browser compatibility should be tested regularly, followed by program adaptations if needed. Second, minimum requirements on type and version of operating system and browser software should be communicated to both researcher and participant. And third, browser and operating system detection should be applied in order to determine the type of software the participant is using. Participants may subsequently be informed about problems with the software and actions that can be taken (e.g. to download a new version of the browser). Detected software specifications can be made available to researchers to possibly use as a control variable.

Hardware standards

Participants may use a variety of processors, computer screens, input systems, audio systems, and types of devices (personal computer, laptop, tablet, or smartphone), all of which may affect test performance. Therefore, hardware should be tested on several systems, including older, low-end hardware, such that most of the target population is expected to have a better system. Subsequently, minimum requirements can be communicated.

Processor and memory

Even a computer with high processing power may run slowly if there is little memory available. Therefore, first of all, participants should be instructed to close other applications before starting the test. Second, minimum requirements (processor and RAM) can be determined. And finally, processing speed may be assessed during the start of the online test. This can be done while participants are viewing the welcome screen with information about the assessment: a background

script—invisible to the user—may execute a number of tasks where the total execution time may not exceed a predetermined number of milliseconds. If this time is nonetheless exceeded, this is recorded as "running on a slow computer" and may influence interpretation of the data.

Screen

The resolution, refresh rate, and size of the screen may influence stimulus presentation (Cernich, Brennana, Barker, & Bleiberg, 2007). To minimize this effect, the design of an online test should be plain and clear, and not highly dependent on color. Instructions about the positioning of the screen should be provided to participants.

Input system

Currently, there are many systems available to transmit behavioral input to the computer. Participants can use a mouse (available in a large variety of types and models), track pad, digital pencil, touch screen, etcetera, to complete an online test. Participants' performance is likely to be optimal if they use the input hardware to which they are accustomed. However, it is important to consider the different characteristics of different systems. Plant, Hammond, and Whitehouse (2003) have shown that there is a large variability in input registration between different types and models of mice, generating statistically significant effects in a test on simple visual reaction times. Therefore, the input hardware of choice should always be registered, for example by a (multiple choice) built-in question on type of input device. If timing characteristics of the hardware are known, recorded times can be adjusted when needed (Plant et al., 2003). However, if input system functionality testing (as an extension of the cross-system tests) shows adverse effects of one type of input system (e.g. touchpad), it can be advised not to use this system. Keyboards can vary in the position of the keys and in the position of the board itself (laptop or personal computer). Results are likely most reliable if participants use the type of keyboard they are familiar with.

Audio system

Instructions can inform participants about the need for specific audio hardware such as headphones or loudspeakers. If timing of audio stimuli is crucial, minimum requirements on type of soundcard can be provided also. Furthermore, calibration of the volume, similar as in programs for video calls, should be built in the testing platform.

Device

The impact of the use of mobile devices on data quality should be considered (Germine et al., 2012). Compatibility to mobile devices requires very specific programming. Therefore, the performance of the test should be tested separately for mobile use. Since it is likely that people will try to take the test on a tablet or smartphone, participants should be instructed about whether or not they can use such device to take the test. If the test is not suitable for such devices, it should preferably not be functional on tablets or smartphones at all. Automated detection of unsuitable devices can be programmed into the test, as well as a subsequent warning message ("unsuitable device") to the participant.

Web design and usability

In order for participants to complete an online neuropsychological test independently, the test should be easy to use. This ease of use is addressed by the "usability" of the user interface. The International Organization for Standardization defines usability as follows: 'the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.' (http://www.usabilitynet.org/ tools/r_international.htm#9241-11). For an online test, usability is applicable to both the individual tests and the test environment. According to Nielsen (2012), usability can be determined by the following: (1) ease of learning, (2) response efficiency, (3) memorability, (4) errors, and (5) user satisfaction (http://www.nngroup.com/articles/usability-101-introduction-to-usability/). Graphical user interface (GUI) design standards address the use of features such as color, layout, and screen design (ITC, 2006) and can be used to optimize these five factors. In general, usability benefits from keeping the design as clear as possible in terms of text, colors, and figures, simple and consistent in terms of layout, and intuitive, such that the meaning and location of response options are evident. Table 3 provides a number of features that may contribute to an online test with adequate usability.

Table 3. Design standards for an online neuropsychological test.

Design
• Clear differentiation between test items and instructions
• Consistent layout and location for instructional text and prompts
• Consistent (calm) color settings
• Only relevant information displayed on the screen
• No distracting logo's/ images
• Advanced media features only when justified by validity
• Undo or back buttons where possible
• For textual input: cursor automatically placed in text field
• No scrolling needed
• Questions (in questionnaires) presented one at a time
Instructions
• Clear instructions on use of the system visible whenever appropriate (do not rely on memory)
• Recurring important instructions (e.g.: "as fast as possible")
• Information on what to do in case of common problems
• Instructions to set browser to full-screen mode
Feedback
• Appropriate information on system status (current position in time structure)
• Clear error messages, with suggestion for solutions
• Direct feedback (point out error as soon as is made)
• Clear task instructions (see paragraph "task comprehension")
• Repeatable (video) instructions
• Option to request help

In order develop a test with sufficient usability, it should be pretested on its particular target test population. Usability studies pretest materials and help to improve the usability of online tests (Huff, 2006; Reips, 2002d). Feedback—by observing users who take the test and by employing questionnaires after conducting the test—from small numbers of representative

users should be used to optimize test development.

Data recording

Test scores are derived from the behavioral data recorded during the test. Therefore, it is crucial that test software promotes standardized and adequate recording of behavior. The test should respond quickly to commands in order to register behavior correctly. Timed elements should only include testing time—excluding processing time—and data have to be transferred to a database regularly (for example after the completion of every element) to prevent information loss in case of instable Internet connections. Besides behavioral information, data can include information on computer configuration such as follows: (1) type and version of web browser; (2) type and version of operating system; (3) screen resolution; and (4) accuracy of the computer's timing response (Reips, 2002d). As suggested above, this information can be used to check for minimum requirements, apply corrections based on technical variation, or study the effect of technical variation.

Contextual issues in online neuropsychological testing

Similar to technical variations, there may be several variations in the individual participants' testing context that may increase measurement noise for online unsupervised testing and that researchers employing online neuropsychological assessments may want to control for. These contextual differences may pertain both to the individual (e.g. computer skills and disabilities that may affect participants' responses), as well as to the individuals' physical context while completing the test (e.g. (dis)comfort and distraction).

Computer skills and disabilities

In order to facilitate optimal test performance in face-to-face assessments, it is also the test leader's task to be flexible to the needs of the participant (Lezak et al., 2004). These needs may depend on participants' computer and language skills, and disabilities other than cognitive –such as visual, hearing, motor, or reading impairment. Individual differences in computer familiarity and computer skills may influence how participants interact with the web page, use response devices, or respond to stimuli (Bauer et al., 2012; Parsey & Schmitter-Edgecombe, 2013). In particular, tests which rely heavily on rapid visual scanning, rapid response time, and accurate keyboard use seem to be sensitive to the participant's level of computer familiarity (Iverson et al., 2009). Even though sufficient computer familiarity is becoming more common in current research populations (Chuah, Drasgow, & Roberts, 2006), it is likely that participants will have varying computer skills. In addition, motor or sensory disabilities might also influence how participants complete an online test. These variations should affect cognitive test performance as little as possible. Test programming and design should limit increase in mental processing demands (i.e. cognitive load) in less-skilled or disabled computer users.

Motor problems and computer skills

Besides general GUI standards to increase functional accessibility, there are more specific settings to consider. For participants with motor problems or computer inexperience, the demand on computer skills needs to be low. Participants should be able to give behavioral input with few and simple actions (mouse clicks or key presses). Drag-and-drop actions are a relatively advanced response mode (Hansen et al., 2015) and should therefore be applied with care. Brief instructions on computer use can be provided after an unsuccessful practice trial or upon request (e.g. "you can move the object as follows: (1) place the mouse pointer on the object; (2) press and hold the (left) mouse button; (3) move the mouse to the desired location; and (4) release the mouse button there."). Information on the used hand and the preferred hand can be registered. Furthermore, user information on computer familiarity can be gathered and tests assessing computer skills (mouse and typing skills) can help to interpret and, possibly, adjust scores. However, when adjusting for computer skills and familiarity, it is important to take into account confounding variables such as age and education (Hansen et al., 2015).

Visual disabilities

For participants with visual disabilities, the visual stimuli need to be large and of high contrast. In addition, pretest instructions can advise on the use of glasses and other visual aids. For participants with color blindness, the use of colors should be restricted. Instead of full range colors, different hues of gray can be used to give meaning to elements on the screen.

Hearing disabilities

For participants with hearing disabilities, optional subtitles should be provided along with all spoken text. Furthermore, a sound level calibration can help to set the volume to a comfortable level before the start of the test.

Reading disabilities, language deficiencies, and international use

For participants with reading disabilities or language deficiencies, the test should be language independent for the larger part. Instructions can be represented primarily by figures, screenshots, or animations. Supplementary spoken instructions can be added separately. Aside from making tests more accessible to a variety of participants, such an approach also allows platforms more flexibility in instructional language and thus is ideal to promote international, yet standardized, use (Rohlman, Sizemore, Anger, & Kovera, 1996). In international use, it is important to use good translations (Naglieri et al., 2004) and to test the equivalence of adapted versions (ITC, 2006). The use of language, drawings, content, and graphics that are country or culture specific should be avoided (ITC, 2006).

Because tests should be presented as standardized as possible, we recommend to build in general functions that result in high usability for the majority of participants. Variability in presentation of the web page should be limited. For example, tests should preferably have a standard large font size, instead of enabling an adjustable zoom factor.

In the case of complete lack of computer skills or serious disabilities in participants, online testing may not be feasible. In such cases, face-to-face assessments are preferred (Bauer et al., 2012). The test manual should indicate which participant characteristics lead to exclusion for which specific online test. Most likely, this can be determined on the basis of a few basic questions on capabilities and disabilities. Milder disabilities or slight lack of skills are, however, up to interpretation of the researcher.

Environmental control

The test conditions in a participants' home environment must be sufficiently controlled in order to obtain reliable data. Standards applying to assessments in the laboratory setting should be emulated as much as possible, and the testing environment needs to promote a focus on optimal performance. To accomplish this, clear general instructions can be provided (ITC, 2006; Tippins et al., 2006). These general instructions should address all issues a test leader would normally point out before the assessment: user expectations (type of tests and time schedule), how to create an optimal working environment (comfortable seating, clean work-surface), how to limit possible distractors (being alone in the room, turn of phone and television, use headphones), and the type of mind-set expected from the user (optimal motivation and concentration). The unmonitored setting requires stressing that under no circumstances are participants supposed to receive help, or use paper and pencil, while completing the test. In some cases, an identity check can be used to assure that the participant is who he/she purports to be (Bauer et al., 2012). As a low-technology solution, participants can be given a personal password that can only be used once (Reips, 2002d). More secure high-technology solutions, such as retinal scans or fingerprints, are currently not feasible for large-scale testing (Rovai, 2000). Identity checks by online video surveillance may be feasible, since most modern laptops have built-in webcams and external webcams are inexpensive. Such an identity check can be valuable in clinical settings where test results from a specific patient need to be gathered. However, in clinical research studies, participants are often highly motivated and intrinsically driven to complete the online test themselves (Bilder, 2011), which makes high-technology solutions unnecessary. A study of Wefel and colleagues (2014) showed that in a dataset of a total of 534 breast cancer patients and 214 healthy controls who completed neuropsychological tests and embedded performance validity tests, only 1 patient met the criterion of non-credible performance (Wefel et al., 2014).

Psychological issues in online neuropsychological testing

One of the important functions of a test leader in face-to-face neuropsychological assessments is to take care of participants' psychological needs in order to optimize the focus (motivation and concentration) needed to obtain a reliable measure of their cognitive functioning. There are many psychological factors that may influence participants' focus, and in the unmonitored setting, signs of motivational or mental status issues can easily be missed (Bauer et al., 2012). Therefore, to prevent participants' online tests performance from being hampered by psychological factors, it is important to create an online environment that attends to participants' psychological needs. We will discuss four main components: task comprehension, task involvement, fatigue, and stress and anxiety.

Task comprehension

Task comprehension is crucial for optimal task performance. If participants fully understand how to complete a test, they can focus on optimal performance. In contrast, task ambiguities could lead to unnecessary low test scores. It is therefore important to train participants well before the actual test is presented. This can be done by means of instructions and practice trials with feedback (Rohlman et al., 1996). In face-to-face assessments, the test leader checks whether instructions have been understood and acts to improve comprehension when needed and where possible. In the unmonitored setting, such continuous monitoring of participants' task comprehension is more difficult. However, if instructions and practice with feedback are presented properly, task comprehension can be optimized. To do so, it is important to consider participants' cognitive load at all times. The cognitive load theory (Chandler & Sweller, 1991) proposes that cognitive load should be kept within working memory capacity to enable learning (Feinberg & Murphy, 2000). This means that cognitive load required to process the instructions and additional feedback should be kept low (Brünken & Plass, 2003) – not only to optimize task comprehension but also to subsequently optimize task performance.

In order to keep cognitive load low, instructions should be clear, but also concise and engaging. Written instructions are often skipped or not fully attended to by users of computerized tests (Rohlman et al., 1996). Reading plain text on the screen requires skills and effort; new information is processed linearly and learned slowly (van Hooijdonk & Krahmer, 2008). Too high a demand on working memory causes participants to easily lose track and even to stop reading. In contrast, audio-visual presentation of instructions promotes the use of separate memory systems during processing: dual processing. This could prevent cognitive overload (van Hooijdonk & Krahmer, 2008) and enhance learning as individual capacity is optimized (Brünken & Plass, 2003). Several studies have shown the advantage of dual processing in the effectiveness of instructions (e.g. (Brünken & Plass, 2003; Mayer & Anderson, 1991). Video combined with verbal instructions potentially enhances motivation (Choi & Johnson, 2005) and was shown to be a preferred method for learning (van Hooijdonk & Krahmer, 2008). Additional advantages of video instructions are that they provide the opportunity to show real footage (screen captures) and a good representation of temporal aspects of the test (van Hooijdonk & Krahmer, 2008). Furthermore, video instructions are standardized by nature, while, in contrast, participants may vary widely in the time they spend on processing written instructions. When used, verbal instructions should be carefully scripted and not interfere with the images, to prevent unnecessary cognitive load (Mayer & Moreno, 2002). However, video instructions might be skipped or disregarded. therefore, checks for understanding of instructions may be a useful addition, for example by using comprehension questions that need to be answered correctly before proceeding with the actual test (Crump et al., 2013).

After instructions, a practice session with feedback can further check for comprehension and prepare participants for the actual test. During practice, a short version of the test is completed at an undemanding level. Behavior from the practice session can be used to generate context-sensitive feedback, e.g.: "try to respond as fast as possible". Optimal feedback depends on

the user and his or her computer experience (Jacko et al., 2004). In general, optimal feedback is meaningful, concise, non-alarming, clearly visible and/or audible, and adaptive to the participant's performance level (Hodges, 2004; ITC, 2006). Furthermore, in cognitive demanding tasks, immediate feedback proves to be more effective in optimizing task performance than delayed feedback (Kulik & Kulik, 1988). In some cases, however, practice sessions are not possible, because they interfere with testing the construct of interest. For example, practice of a word learning test is not advisable, as practice with words of the actual test would improve test performance, and practice with non-test words could cause intrusive memories during the actual test.

Task involvement

The effort one puts into a test is very likely to be influenced by a participants' task involvement: the extent to which optimal performance is the participant's own goal. Task involvement depends heavily on the trustworthiness of the setting (Reips, 2002d), and on the personal importance of the assessment (Hodges, 2004). Reips (2000) mentions the following general credibility-enhancing elements: (1) provide the name of your institution; (2) emphasize the scientific purpose of the test; (3) ensure (and keep) confidentiality; and (4) provide contact information. For clinical populations, the overall appearance of the test should not be too impersonal or medical, because this could cause stress and negative associations with being ill. The importance of the assessment can be emphasized in the pre-assessment communication (e.g. during recruitment) by explaining the context and personal relevance. Emphasizing the importance of participation in the research study is also likely to induce engagement during the test. This can be done, first, by thanking participants for participating before the test; second, by providing participants sufficient communication opportunities, and freedom for additional input during questionnaires; and third, by giving participants personalized feedback which confirms that answers have been processed (e.g. by a pop-up text or visual/auditory animation). Finally, feedback on task performance may be included as a motivational cue (Bangert-Drowns, Kulik, Kulik, & Morgan, 2012; Economides, 2005). The wording of such feedback should be carefully chosen in order to positively affect participants' motivation. Negatively framed feedback ("80% of participants scored better") can discourage participants; positive framing ("you are in the top 80%") is generally considered more motivating. Also, unspecific or generally phrased feedback ("well done!") can irritate and demotivate participants. Even though most participants of clinical research studies are highly motivated, dropout will occur. Placing motivationally adverse factors (e.g. asking for sensitive personal information or specifying test duration) mainly at the start of the assessment may reduce the dropout ratio (Reips, 2002d).

Fatigue

Studies so far do not provide a clear consensus on the association between fatigue and performance on neuropsychological tests, but when associated cognitive impairments have been found, they include problems in sustained attention, concentration, reaction time, and processing speed (Lezak et al., 2004). A fatigued patient may be more likely to have problems focusing during the test. Therefore, participants should be well-rested before the start of the assessment and

the assessment itself should not be overly exhausting. General pretest instructions can advise participants to take the test when they feel well-rested. Moreover, for longer assessments, breaks should be included at fixed stages. Offering short rest and refreshment breaks can make completion more comfortable (Bennett-levy, Klein-boonschate, Batchelor, Mccarter, & Walton, 2007). These breaks provide the opportunity to relax from the need to concentrate and perform.

In online assessments, it might be a good idea to standardize breaks both in terms of length and type of activity. If one participant goes to make coffee, while another spends the break answering e-mails or texts from friends, the effect on performance after the break is likely to be diverse. A structured, timed, and not too demanding break activity enables participants to relax in a controlled manner. This can be done, for example, by including short leisurely videos after completion of a set number of tests as an integral and standardized part of the test battery. In offering standardized entertainment material as a break activity, it is important to consider the possible influence that it may have on testing performance. For example, the use of test elements (such as specific words) can interfere with responses on the test. Breaks should be well-timed in order to optimize the motivation needed for different tests. If particular tests are very demanding, breaks could be most appropriate prior to or after such tests. Participants should be informed about the regulated breaks before starting the test (ITC, 2006). In addition, “next-buttons” can further pace the assessments by providing the opportunity to continue to a next element when ready (while being encouraged to not take long breaks in between tests). A clear and quiet test design can prevent additional fatiguing. Finally, in order to gather more information on participants’ general and current level of fatigue, a fatigue questionnaire can be included in the test battery.

Stress and anxiety

Feelings of stress and anxiety increase cognitive load (Buckelew & Hannay, 1986)—just like high cognitive load can increase feelings of stress and anxiety (Brünken & Plass, 2003)—and, furthermore, lead to an uncomfortable assessment. Therefore, the amount of stress and anxiety should be limited as much as possible. There are a number of factors that can cause or reinforce stress and anxiety in participants during the test:

Pressure to perform

A moderate amount of performance pressure can facilitate optimal performance, but too much pressure is likely to reduce test performance. Fear of impairments, fear of appearing foolish, and a threatening test situation may prevent optimal performance (Lezak et al., 2004). Therefore, pretest instructions should emphasize the importance of “trying your best” but at the same time reassure participants that some test are hard and that everyone will make mistakes. If participants know what to expect—by explaining how to prepare for the test and how much time and effort it will take—they will be more at ease. During the test, there should be information on test progression, and the option to request for help. Feedback on performance needs to be stress-reducing: useful information in a friendly tone.

Computer anxiety

Another stress-inducing factor may be computer anxiety. Computer attitudes can attenuate the influence of computer experience (Browndyke et al., 2002; Fazeli, Ross, Vance, & Ball, 2013; Mahar, Henderson, & Deane, 1997), but especially participants with little computer experience can experience stress when performing computer tests (Mahar et al., 1997). This is an additional reason to test participants on basic computer skills and to keep the demands on computer skills as low as possible. General instructions can emphasize these low demands. Furthermore, a few questions on computer attitudes can be included in order to monitor levels of computer anxiety (Browndyke et al., 2002). This will be less of an issue as nowadays people of all ages are more likely to have computer experience (Wild et al., 2012), but should nevertheless be a point of attention.

General anxiety level

The general anxiety level of a participant could also influence stress during assessment and in turn effort during assessment and performance (Eysenck & Calvo, 1992). High trait anxiety or experienced life stress is found to reduce measures of working memory capacity (Derakshan & Eysenck, 1998). Participants could have difficulties concentrating on the assessment or feel the need to perform especially well. A questionnaire on current stress and anxiety levels can be included in order to be informed about clinically meaningful elevated levels.

Priming

Priming of implicit memory can lead to underachievement in certain subgroups of participants (Schmader & Johns, 2003). For example, it has been shown that priming participants with a disability stereotype can lead to decreased manual dexterity and slowed performance (Ginsberg, Rohmer, & Louvet, 2012). The concern to confirm a certain stereotype increases cognitive load and subsequently affects cognitive performance. It is therefore best to steer clear of all statements on expected performance, and to ask questions on priming-sensitive topics such as (computer) anxiety, illness, neuropsychological complaints, gender, age, race, or level of education not before but after the neuropsychological assessment.

Privacy concerns

Concerns about the security and privacy features of the Internet may influence performance (Naglieri et al., 2004) and, in addition, can lead to dropout (Reips, 2002d). Therefore, measures that are taken to securely transfer and save data—such as encrypted transfer and separate storage of personal and performance data—should be clearly communicated to participants, or made available on request.

Quality assurance

Even though, as outlined above, the adverse influence of technical, contextual, and psychological factors may be minimized, differences in online and offline test results may still emerge, stemming, for example, from differential psychometric properties. In this paragraph, we will discuss how to

collect information on psychometric properties, and the necessity to collect up-to-date norm scores, provide updated instructions for professional users, as well as technical maintenance in order to ensure optimal data quality from online neuropsychological assessments.

Psychometric properties

Even if online tests are based on well-studied paper-and-pencil or computerized tests, it cannot simply be assumed that they have similar psychometric properties as their offline counterparts (Barak & English, 2002; Bauer et al., 2012; Chuah et al., 2006). In online tests, many test features—such as input mode, stimulus presentation, feedback, timing, etcetera—need to be adapted in order to establish high usability and facilitate optimal performance. Therefore, online tests based on established tests are by definition different and should be considered as new tests. This means that the interpretation of their results requires re-establishing reliability, validity, and norm scores. If diagnostic classification is required, information on sensitivity, specificity, positive predictive power, and negative predictive power should be considered also (Bauer et al., 2012). Before deployment of the test, psychometric properties should be evaluated under conditions representative of future test conditions and populations (ITC, 2006). After deployment of the test, psychometric properties should be re-assessed in the case of any alterations to the test or the testing environment, or application in another target population. Currently, there are no conventional benchmarks for psychometric properties of (online) neuropsychological tests. The allocation of benchmarks for reliability and validity is somewhat arbitrary, but adherence to standardized psychometric criteria could improve validity and test application (Gates & Kochan, 2015; Wild et al., 2008). Studies on reliability and validity should determine minimum standards on beforehand.

Reliability

Reliability refers to 'the degree to which a test is free from measurement error' (de Vet, Terwee, Mokkink, & Knol, 2011). A reliable instrument produces consistent scores over time, alternate forms, or across raters, as long as there are no changes in the measurement construct. Reliability of online neuropsychological tests will most likely be evaluated by assessing test-retest reliability, in which relative consistency over time is determined. This relative consistency is best quantified by intraclass correlation coefficients (ICCs) (Weir, 2005). An ICC of .80 indicates that an estimated 80% of the observed score variance comes from "true variance". However, even though criteria have been suggested, there is no consensus on benchmark ICC values. First, criteria are always to some extent arbitrary. Second, the magnitude of the ICC varies depending on the type of ICC selected and the variability in the data. It is therefore important to take the type of ICC selected into account when determining minimum standards. In addition, data from (within subject) repeated testing can be used to evaluate general levels of measurement error (Standard Error of Measurement) (Weir, 2005). Taking these factors into account, data on reliability can be used to optimally interpret test scores.

Validity

Validity refers to 'the degree in which a test truly measures the construct(s) it aims to measure' (de Vet et al., 2011). Validity is a property of the interpretation of test scores within a specific context, not a property of the test itself. Therefore, validity studies should be conducted under appropriate conditions. Concurrent validity of a new online neuropsychological test can be evaluated by comparing its performance scores with those of a "gold standard" test, most likely the (well-studied) paper-and-pencil or computerized test on which the online test was based. Test scores of the new test should correlate with those of a gold standard test to a specific degree, which may be set beforehand. In doing this, it is important to take test-retest reliabilities of the gold standard, and possibly also of the new test, into account. For example, if a gold standard test has a test-retest correlation coefficient of .70, the correlation between the gold standard and its online equivalent cannot exceed .70, and will in all likelihood be lower. If the reliability of the new tests has been studied also, the correlation with the gold standard cannot exceed *√(reliability[gold standard] x reliability[new test])* (de Vet et al., 2011).

If a new test clearly does not meet preset criteria, it is likely that it does not or not merely measure the construct it aims to measure. Then, test performance could also reflect, for example, computer skills. In such cases, adaptations of the test, the test environment, or scoring procedure are required, followed by new validation studies.

Norm scores

An individual's neuropsychological test scores can be interpreted by comparing them to those of a normative sample. Online neuropsychological tests require online-specific norm data. Normative samples may be groups of healthy people, clinical samples, or other specified groups (e.g. elderly, or inhabitants of a certain country or region). To facilitate a correct interpretation of test scores, normative data should be representative of the test population and derived under similar circumstances. Based on analyses of variance, norm scores are often stratified by population characteristics such as age, gender, and level of education (Strauss, Sherman, & Spreen, 2006). In addition, for online testing, computer skills can be taken into account. Since online testing is well-suited for continuous gathering of normative data, up-to-date norm scores can be made available for interpretation with relative ease.

Instructions for professionals

For the use of neuropsychological tests, professionals (researchers, clinical neuropsychologists, psychological assistants, etcetera) need to be informed about when best to use them, how to use them, and how to interpret their results. This should include information on when it is not feasible to use the test and how to interpreted test results in case of influential technical specifications or participant characteristics. Table 4 describes which issues a user manual should address.

Table 4. Recommended content of the user manual of an online neuropsychological test battery.

Measurement construct	• Which construct is measured by each individual test
Test development	• Developmental decisions and pilot studies
Administration	• Optimal test conditions
	• Minimal hardware and software requirements
	• Suitable devices
	• Participant characteristics that make online testing unfeasible (e.g. lack of computer skills)
	• Instructions as presented to the test takers
Scoring	• Data registered
	• Test scores calculated
	• Compound scores calculated
Portal use	• Communication with participants
	• Access to database
Psychometric properties	• Validity
	• Reliability
	• Normative data
	• Information on how motor or sensory impairment may affect performance
	• Information on how technical specifications may affect performance
Technical support	• Technical support for professional users
	• Technical support for test takers (in case of technical problems during the assessment)
	• What to do in case of common technical problems (frequently asked questions)
	• Frequency and scope of software updates and backup of collected data
Information on data security	• Where the data is stored (server)
	• Who has access to the data
	• Which measures are taken to secure patient privacy

Technical maintenance

In order to keep performance of the test consistent, regular checks on compatibility with third party hardware and software should be performed, followed by updates if required. Since technical problems can lead to poor data quality and dropout, technical maintenance should be continuous. This requires good documentation on current technical specifications and adaptations thereof (Schlegel & Gilliland, 2007). Systematic checks should be planned regularly and can be automated (automatically performed by aforementioned online services). This way, influence of changing technical characteristics can be kept to a minimum. Furthermore, to ensure that security of the collected data is maintained, first, security certificates should be updated regularly, and second, separate servers for storage of test scores and participant data should meet the latest safety requirements.

DISCUSSION

Recent developments in research on cognitive (side)effects, treatment, and diagnosis call for online neuropsychological measurement tools that allow for large-scale data collection. The main advantages of these online tools are as follows: standardization (by computerization), accessibility, and efficiency. Currently, few online neuropsychological test batteries are available; tests that are available often still require on-site supervision by a test leader, or lack proper psychometric evaluation. In addition, there are no guidelines available specific to online neuropsychological testing in the unmonitored setting. The unmonitored setting does, however, require specific attention since it may add noise and reduce reliability of cognitive data. In this paper, we have outlined achievable prerequisites for the proper development and use of online neuropsychological tests, with the focus on reliable measurement of cognitive function for clinical studies in an unmonitored setting: how to use computerization beneficially and, at the same time, minimize added noise from the unmonitored setting.

Data from neuropsychological tests should represent optimal performance, with minimal influence from technical, contextual, and psychological participant variation. To yield standardized data collection under a variety of technical conditions, the following *technical* factors can be beneficial: deliberate client-side/server-side processing, insensitivity to hardware and software differences (e.g. processing power or web browser), monitored computer configurations, and smart data recording. Concerning *contextual* factors, test design can be made suitable for patients with impairments or language constraints, and should include clear instructions on appropriate testing environment. Beneficial *psychological* factors are a test design and test instructions that entail little cognitive load and optimize motivation, a trustworthy setting, securely recorded test scores, and avoidance of impeding psychological factors such as fatigue and anxiety. In addition, correct interpretation of online data and long-term data quality is crucial for the functionality of an online neuropsychological test. Online adaptations of existing neuropsychological tests should be regarded as new tests, and therefore, test properties, such as reliability, validity, and norm scores, need to be newly determined. Finally, maintenance of the test itself, its norm scores, and data safety need to be ensured.

Complete standardization of unmonitored online testing is not feasible. Above all, factors that may add noise to cognitive data should be prevented by smart test development. Second, cross-system/browser tests and usability studies should be used to determine minimum requirements. Third, clear instructions on the test setting should be provided to participants and researchers. And finally, an end user manual should provide instructions on how to interpret test scores. Noise factors that cannot be ruled out or cause noise to a lesser extent can be reported so that they can be taken into account. It is important to realize when the use of an unmonitored online test is not feasible or will lead to poor reliability. In some cases, it will be advisable not to use online testing; the manual should guide this decisional process. For example, it will not be feasible to use an unmonitored online test in cases of outdated hardware or software, complete lack of computer skills, or serious disabilities. Furthermore, participants will need to approach the test with sufficiently serious attitude in order to rely on them taking the test themselves, in a fair and

motivated manner.

This paper has discussed the many factors involved in the development and use of unmonitored online neuropsychological tests. Note that, in our view, all the steps mentioned (from development, to maintenance) should be taken in order to enable reliable online cognitive data collection. This indicates that the development and proper use of high-quality online neuropsychological tests is quite complex. Obviously, a collaborative effort among developers to create and maintain appropriate measurement tools would facilitate this process. Neuropsychologists and developers could work in a modular fashion to develop such measurement tools, for example by using online collaborative workspaces such as GitHub. Additional benefits of such collaborations may lie in (1) transparency and members' involvement with the measurement tools, properties, and applications; (2) keeping tests, surrounding technology, and normative data up-to-date; and (3) widespread and large-scale test application. The latter would facilitate the comparability of study results, data pooling, and improving and specifying normative data, e.g. applying to different countries, patient demographics, and languages. Currently, online neuropsychological tests may be under development by research groups all over the world for use in their own specific studies, but such technology is hardly made available for general use. Tests developed through collaborative efforts might, at first, focus on creating generic test platforms, addressing basic measurement constructs and predominantly language independent tests. This will ensure immediate benefits for all involved. In later stages, groups can adapt tests, or create new, more specialized, tests, based on the technology and code developed in the earlier stages. It would be beneficial to create tools that can be widely deployed and enable collaborative large-scale online cognitive data collection.

Concerns formulated with regard to online neuropsychological testing are a good starting point to develop and use online measurement tools properly. Many of these concerns are applicable to face-to-face testing as well: tests should be carefully developed, evaluated, and maintained; conditions under which tests are taken need to be monitored and, as much as possible, controlled; tests should measure optimal performance with as little interference by stress or anxiety as possible; and test metrics should be made available for appropriate use and interpretation. Currently, no online neuropsychological test platforms are available that properly address all these concerns. However, there is a clear need for online data collection. Investment in the development of online neuropsychological test platforms is now desired. With this paper, we aim to contribute to these developments. In the future, in some research or clinical contexts, online assessments may replace or complement face-to-face assessments or act as screening tools. Thorough psychometric studies will be needed in order to decide whether such online tests can indeed be used as measurement tools in neuropsychological research. Future use of online neuropsychological tests will depend on these psychometric properties, along with the general strengths and weaknesses of the online methodology within specific research populations.