Moving psychological assessment out of the controlled laboratory setting: Practical challenges

Terje B. Holmlund

University of Tromsø

Peter W. Foltz

University of Colorado & Pearson

Alex S. Cohen

Louisiana State University

Håvard D. Johansen

University of Tromsø

Randi Sigurdsen

University of Tromsø

Pål Fugelli and Dagfinn Bergsager

University of Oslo

Jian Cheng, Jared Bernstein and Elizabeth Rosenfeld

Analytic Measures Inc, Palo Alto, California

Brita Elvevåg

University of Tromsø

Norwegian Centre for eHealth Research

Author Note

Terje B. Holmlund, Department of Clinical Medicine, University of Tromsø;

Peter W. Foltz, Institute of Cognitive Science, University of Colorado and Pearson;

Alex S. Cohen, Department of Psychology, Louisiana State University;

Håvard D. Johansen, Department of Computer Science, University of Tromsø;

Randi Sigurdsen, Faculty of Law, University of Tromsø;

Pål Fugelli, University Center for Information Technology, University of Oslo;

Dagfinn Bergsager, University Center for Information Technology, University of Oslo;

Jian Cheng, Analytic Measures Inc, Palo Alto, California;

Jared Bernstein, Analytic Measures Inc, Palo Alto, California;

Elizabeth Rosenfeld, Analytic Measures Inc, Palo Alto, California;

Brita Elvevåg, Department of Clinical Medicine, University of Tromsø, Norway and Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

This research was funded by awards from the Research Council of Norway to Brita Elvevåg (#231395) and to Håvard D. Johansen (#263248).

The data presented in this paper have not been previously published or publically presented.

Correspondence concerning this article should be addressed to Terje B. Holmlund, Department of Clinical Medicine, UNN Åsgård, Postbox 6124, 9291 Tromsø, Norway. Email: terje.holmlund@uit.no

Abstract

Behavioral assessment using smart devices affords novel methods, notably remote self-administration by the individuals themselves. However, this new approach requires navigating complex legal and technical terrain. Given the limited empirical data that currently exists, we provide and discuss anecdotes of the methodological, technical, legal and cultural issues associated with an implementation in both US and European settings of a mobile software application for regular psychological monitoring purposes. The tasks required participants to listen, watch, speak, and touch to interact with the smart device, thus assessing cognition, motor skill, and language. Four major findings merit mention: First, moving assessment out of the hands of a trained investigator necessitates excellent usability engineering, such that the tool is easily usable by the participant and the resulting data relevant to the investigator. Second, remote assessment requires that the data are transferred safely back to the investigator, and that risk of compromising participant confidentiality are minimized. Third, frequent data collection over long periods of time is associated with a possibility that participants may choose to withdraw consent for participation thus requiring data retraction. Fourth, data collection and analysis across international borders creates new challenges and new opportunities because of important cultural and language issues that may inform the underlying behavioral constructs of interest. In conclusion, the new technological frameworks provide unprecedented opportunities for remote self-administered behavioral assessments but will be most productive in multidisciplinary teams to ensure the highest level of user satisfaction and data quality, and to guarantee the highest level of data protection.

KEY WORDS: mobile health, mental health, assessment, information security, consent

Public Significance Statement

Technologically it is now possible to collect a variety of behavioral measurements frequently and remotely via smart devices operated by the individuals themselves. This paper describes the implementation in the United States and Europe of mobile psychological assessment software for regular monitoring purposes. Although new technology affords an unprecedented opportunity for psychological assessment, these new approaches are accompanied by important methodological, technical, legal and cultural issues that must be addressed so as to guarantee the highest possible quality, value and security of each participant's data.

Introduction

Psychological assessment is critical for understanding a broad range of human behavior, including cognitive abilities and clinical symptoms. Historically, psychological assessment has been conducted by triangulating information from different sources, including behavioral observations that are interpreted by trained experts, historical information about individuals, test performance and self-reports from interview-based tools. Building on this mature tradition, it is now viable to collect a variety of behavioral measurements via mobile devices - such as smartphones - and do so frequently thus providing improvements in both the temporal resolution of measurements (Cohen et al., 2018a) and in the ecological validity since data can be collected remotely via devices operated by the individuals themselves (Trull & Ebner-Priemer, 2013). These technological advances hold promise of being the catalyst for much-needed discoveries in neuropsychology in general (Bilder, 2011), and especially in neurodegenerative conditions (Au, Piers & Devine, 2017) and complex cases with comorbidity such as in epilepsy (Moore, Swendsen & Depp, 2017; Witt et al., 2013). Beyond the mere promise, methods of real-time assessment have provided rich data in the investigation of mood disorders (Ebner-Priemer & Trull, 2009), and been demonstrated to be more precise than retrospective methods of assessing health related behaviour (e.g. in evaluations of cigarette usage (Shiffman, 2009)).

While there is good reason to be optimistic about the future use of digital technology in psychological assessment, there are numerous sobering indications that the actual adoption of computerized techniques in psychological assessment remains limited (e.g., Rabin et al., 2014). Naturally, all psychological assessment tools must be safe and effective and provide information on test reliability, validity, accuracy, and utility (Bauer et al., 2012). This is not a minor feat as acquiring knowledge on operating characteristics such as sensitivity and specificity requires extensive data collection with participants over a wide range of

demographic and diagnostic categories. Clearly the practical challenges involved in obtaining sufficient data in the first instance, so as to be able to make claims on vital test characteristics, can be the roadblocks to widespread adoption. Additionally, what is technically possible can nonetheless be practically infeasible due to the plethora of novel legal challenges that can seem intractable to many investigators.

The purpose of this paper is to discuss the practical challenges and solutions to developing and implementing this technology for moving assessment out of the hands of a trained investigator and a controlled laboratory setting to self-administration remotely by the individual themselves. Solving these practical challenges will facilitate the acquisition of knowledge on safety and utility required for the trust of practitioners and patients, ultimately enabling a future with wide-spread use of digital technology for psychological assessment.

We discuss methodological, technical, legal and cultural issues, highlighting practical lessons learned through the development and implementation of a mobile software application for remote, frequent and self-administered psychological assessment in both the United States and in Norway. While psychological assessment can span a wide range of domains, from clinical and consulting practices to educational and organizational psychology (Ben-Porath, 2016), our software application focused on assessment of cognitive functions and affective states for the purpose of detecting clinically relevant change in patients with severe mental illness. Therefore, assessment items were similar in form and structure to standardly employed neuropsychological tests, but were designed so that they could be remotely self-administered daily (see Figure 1). The items selected contained tasks that required speaking (story recall, picture descriptions, a modified Stroop task, category verbal fluency), performing touch-screen responses (modified trail making test, modified digit span task, spatial recall span, serial letter recall, synchronization-continuation finger tapping task) and self-report on subjective mental states using on-screen sliders (with questions such as “How

happy are you today?", designed to assay negative and positive affective states). These assessment tasks were then built into software that could run on the iOS platform (a mobile operating system created and developed by Apple Inc.), which has the advantage of being easy to deploy to smart devices such as smartphones and low-cost internet-connected iPods. The software application was made available for download from the Apple software store, providing a framework that permits fast development of highly usable interface components including video and graphic display, speech recording, and capture of gestures and actions.

[Figure 1 about here]

While many mobile assessment tools have been in development in recent years (e.g., Allard et al., 2014; Brouillette et al., 2013; Frings et al., 2008; Jongstra et al., 2017; Kennedy et al., 2011; Riediger et al., 2014; Schuster, 2015; Schweitzer et al., 2017; Sliwinski et al., 2016; Tieges et al., 2015; Timmers et al., 2014; Tiplady, Oshinowo, Thomson, & Drummond, 2009) the current software had the additional challenge to implement tools specifically using speech processing, and do so across two different languages, and thus also within different cultural and legal settings. A total of 353 participants used the software application over three data collection trials. Of these, 219 were healthy volunteers and 134 were patients with a range of diagnoses of psychosis spectrum disorders, substance abuse disorders, and affective disorders. Two of the trials occurred in the United States and one in Norway, and all trials were approved by the local Research Ethics Committees (LSU Institutional Review Board, #3618, Norwegian Regional Ethics Committee, REK nord, #2014/85). Overall, this approach and tool was well accepted by the participants and the data collected were robust. To achieve and establish this, two surveys were conducted, one before implementation in order to guide development and one during the data collection proper. Details of some of these findings are expanded on in the discussion of methodological challenges below.

First, we report on the importance of solid usability engineering of the software to ensure utility and acceptability, as well as providing data that are comparable, but not necessarily identical, to traditional testing methods. Second, we address the fundamental issue that remote assessment requires that data are somehow transferred back to the investigator. This transfer can either be done manually, by physically transporting data on portable storage devices, or by transferring data over the internet. We discuss the advantages and disadvantages of these different technical approaches by focusing on safety both in terms of minimizing data loss or corruption (i.e., data integrity) as well as compromise of confidentiality of the participant. Third, daily data collection over long periods of time increases the probability that participants may wish to withdraw their consent for participation and thus retract their data, and so we discuss the technical viability of the notion that participants can have all their data deleted on request. Fourth, we discuss how implementations of these technologies that enable research across international borders thereby create new challenges, new opportunities and new knowledge because of cultural issues - notably language ones - that differentially relate to and affect the underlying behavioral constructs of interest in assessment.

Methodological challenges

The methodological viability of moving assessment out of the hands of a trained investigator and a controlled laboratory to self-administration remotely by the individual themselves necessitates establishing two things, namely that (i) the usability engineering of the device is of such a high standard such that it is acceptable by the user, and (ii) the design of the behavioral tasks is sufficiently robust and constrained in how the tasks can be taken such that the data are comparable, if not superior, to traditional testing methods.

Usability of tool

For an assessment tool to be adopted it must be considered useful and acceptable by both clinicians and patients. While computerized assessment methodology has been available for many years, the adoption into clinical practice has been modest (Rabin et al., 2014). Therefore, prior to developing the psychological assessment tool, we assessed user needs in a survey format in clinicians (N=90). A core principle from usability engineering was followed of asking users what they actually *did*, rather than what they might *want* (Nielsen, 1993). The purpose of this survey was to establish what current practice is to the assessment type of interest (in our case, psychiatric risk assessment) and what clinicians considered might improve current methods (for details on the user-needs survey, see Cohen et al., 2018b). The clinicians predominantly worked in an outpatient clinic setting (45%) and were trained in psychiatry (19%), counseling (17%) or clinical psychology (16%). The average age was 45 years (SD = 15), a characteristic worth noting as age has been shown to be related to the adoption of new assessment methods (Rabin et al., 2014). For the purpose of designing a general psychological assessment tool, this sample of experts was considered sufficient. However, in cases where more specialized tools are to be developed, for example to be used in randomized controlled trials, other user groups such as patients and their families could - and probably should - be surveyed also. However, given the nature of the illness that was our focus (i.e., serious mental illness) this would have necessitated a very different type of survey for the different user groups.

The information from this user-needs assessment survey informed and constrained the subsequent development of the assessment tool. While there was high variability in the types of measures that clinicians endorsed for assessing risk, there were commonalities in general classes of assessment types used. Since these broadly fell into the categories of cognition, motor skill and language, we developed behavioral assessment tasks that assessed these

domains. Our usability engineering focus was that the tasks and resulting software should be easy and pleasant to use such that it would be acceptable to the participants (Nielsen, 1994), and that the data collection would be efficient and sufficiently constrained in how the tasks could be taken such that the data would be comparable to traditional testing methods. After the tasks had been designed and agreed upon by the various domain experts in the group, a detailed quality assurance process was initiated, finally resulting in data collection in patients and healthy volunteers.

For any approach to be successful, the tools need to be both acceptable to the user, as well as provide the intended information, in this case behavioral assessment. Users have high expectations from mobile tools because of widespread daily use of popular software applications that have an excellent user experience, and so it is necessary that applications are perceived as providing value in return for time investment (Anderson, Burford & Emmerton, 2016; Yang, Maher, & Conroy, 2015). In parallel with this, perceived and actual utility to researchers must be optimized; the data must be available and interpretable without generating an information overflow situation.

Since the initial usability engineering efforts were based on feedback from clinicians and participants in the US, we specifically sought to evaluate the cross-cultural *acceptability* and *appropriateness* of our software application in a Norwegian sub-sample (N = 24, N = 10 female, N = 10 were healthcare professionals, N = 7 were patients receiving psychiatric care, N = 7 were healthy volunteers). The main outcomes were whether or not users liked the application and its overall duration. This was assessed by answering questions using an on-screen slider. The acceptability of the application was rated as good (average of 77.0 on a 0-100 scale; SD 16.3), with the main complaint being that the session durations of about 12 minutes (average of 12.2; SD 2.2) were too long (indicated by a third of the sample). Interestingly, a duration of twelve minutes is about three times longer than what has

previously been reported in comparable tools (range 40s-4 minutes; Moore et al. 2017). The optimal duration was suggested by the sub-sample to be 9 minutes (average of 8.8; SD 4.1). Many participants expressed a preference for an *even shorter* testing session despite our efforts to make it as short as is scientifically meaningful. This latter issue points to a possible limitation of frequent and self-administered psychological assessment where there is no external reward (i.e., an encouraging experimenter physically present) and thus may necessitate other reward incentives such as money (via micro-transactions) or useful insights into participants' own health (via structured feedback about their own responses and performance). Such reward mechanisms will require an infrastructure development to handle transactions and information flow in a compliant manner, but may provide research groups with increased adherence to protocols and a more robust way to acquire behavioral data.

Data from digitalized tasks: The case of the Stroop Color and Word Test

It is possible to transform traditional pen-and-paper tasks into a digital format, thus making administration of the test easier and potentially improving – or automating – the conversion of task behavior into meaningful scores. When transforming traditional paper and pencil tests to a digital format it is useful, and for some tests essential, to establish that the face validity of the tasks is comparable to traditional versions, as digitalized versions are unlikely to be merely a new format, but very likely an entirely new task (Bauer et al., 2012). However, in the case of some tasks, a new approach also brings with it the potential of collecting additional information and for some spoken tasks even a new and automated method of scoring. We illustrate this by discussing the Stroop test as implemented in our software application, as it serves to demonstrate how a classic task can be adapted to a digital format with usability in mind. This task is also well-suited to illustrate the opportunities for

automated scoring of language tasks using state-of-the-art automated speech recognition, as well as the challenges of making stimulus sets comparable across languages.

The Stroop Color and Word Test (Stroop, 1935) is widely used to derive measures of attentional control and has resulted in a massive literature (MacLeod, 1991). In the Stroop task, words, including color names, are presented in various ink colors. The task is to either read the word (e.g., '**YELLOW**' printed in blue ink should be read as 'yellow' and the blue ink font color ignored) or to name the ink color the word is written in and ignore the actual word (e.g., the correct answer to '**GREEN**' printed in red ink is 'red'). Originally, the task was performed by naming colors of words or shapes printed on cards (i.e., the Card Stroop), and a score could be represented by how many responses with correctly named colors were made within a set period (e.g., 45 seconds) timed by a stopwatch (Golden, 1976). This approach has been compared with the more recent computerized Single Trial Stroop approach (Kindt, Bierman, & Brosschot, 1996; Perlstein, Carter, Barch, & Baird, 1998) where stimuli are presented individually on a computer screen, and individual responses made by speaking the color name (timed usually by the voice triggering a voice key), or in the manual version by pressing assigned color buttons (e.g., Waters & Li, 2008). By measuring performance on individual trials within a task it is possible to derive more detailed patterns of performance. However, data from both the Card Stroop and Single Trial Stroop tests have resulted in conflicting results, where the seemingly well-established Stroop Task Interference, represented as slower response times in conditions where there is a conflict between the printed word and its ink color, has not consistently been demonstrated to be disproportionate in clinical populations where this is expected (e.g., in patients with schizophrenia; Westerhausen, Kompus, & Hugdahl, 2011). These findings merit investigation given the widespread adoption of the Stroop task in psychological assessment in general and in clinical populations specifically.

One important way this debate can be resolved is by creating a task that can be used easily by individuals outside of the laboratory setting and do so with a high frequency (e.g., daily) over long periods of time (e.g., months). Therefore, based upon the procedure of Perlstein et al. (1998), we created a variant of a Single Trial Stroop task in our software application (Figure 2 Panel A). Here printed colored words were presented that were either congruent with the ink color (e.g., **RED** in red ink) or were color-word incongruent (e.g., **RED** in green ink), and color neutral words (animal names such as **BEAR**) were also included (Figure 2 Panel B). The task was to name as fast as possible the color that the word was written in. Instructions were presented in a spoken format (male voice) and simultaneously presented in a written – albeit abbreviated – format. The whole task lasted 96 seconds and participants' vocal responses were recorded with the device microphone. Even though the approximately 1.5 minute duration of the task is considerably shorter than many laboratory based paradigms (about one third of the number of stimulus presentations as in Perlstein et al. (1998)), after repeated testing several participants commented that it was "too long" and "boring", making the prospect of increasing analytic power by increasing presentations incommensurable with high acceptability.

[Figure 2 about here]

Our digitalization of this traditional task demonstrates the possibility of going beyond traditional techniques of measuring global performance on tasks, to making it viable to collect and examine individual level performance data (i.e., per response). Additionally, it allows the leveraging of the state-of-the-art automatic speech recognition to automate the accurate timestamping and scoring of verbal responses (Holmlund et al., 2018), thus providing a future assessment framework in which the need for time-consuming (and at times difficult) manual scoring can be eliminated. Such use of mobile technology can potentially create revolutionizing innovations in psychological assessment, and provide critical information

regarding the underlying behavioral and cognitive constructs (Bilder, 2011). Indeed, this approach is extremely promising, notably in terms of the value of the millisecond timing of responses (Dufau et al., 2011) as well as the framework's robustness to possible environmental factors outside of the laboratory (Timmers et al., 2014). Such promise can pave the way for more frequent administration of classic psychological assessment tasks and in a remote fashion.

Technical challenges

Remote assessment can also afford the possibility that it is now controlled by the participants themselves, away from the controlled laboratory environment, a combination that brings with it a variety of technical challenges that several health applications have not addressed (see Huckvale et al., 2015 for a review of 79 applications). Perhaps most notably are the challenges related to transferring data from the devices to the investigator. This can be solved in two distinct ways, namely (i) by manually plugging each device into the investigator's computer and then copying data files onto the research infrastructure, or (ii) instructing the mobile device to send the data automatically over the internet to some online data server. Each method is associated with different logistical and legal terrain, and require mechanisms to safeguard the integrity of data and prevent violation of the confidentiality of participants.

Developing mobile psychological assessment tools for multiple countries simultaneously can present different regulatory challenges. For large scale implementation in the US there may be obligations to adhere to regulations from the Food and Drug Administration (e.g., for computerized cognitive assessment: Title 21 of Code of Federal Regulations, §882.1470 (Neurological Devices, 2017)) or from the Health Insurance Portability and Accountability Act (United States, 2004), while in the EU such

implementation may need to adhere to the respective national regulatory bodies, the EU regulation on medical devices (European Union, 2017) and the General Data Protection Regulation (European Union, 2016). Additionally, even though geographical distance nowadays has little importance when sending and processing digital data, transferring certain types of data across national borders may be illegal if the necessary precautions have not been taken, thus complicating international collaborations. The digitized Stroop task can be used to illustrate some of these 'new' issues: The high quality speech recording of responses in this task can benefit from being processed on powerful servers for automatic speech recognition. To leverage the infrastructure established in our group we would need to transfer European files to US data centers. In the context of our example, unprocessed electronic speech recordings could be considered both personally identifiable and sensitive according to the Norwegian Personal Data Act (for English version, see Datatilsynet (2017)), making transfers to US entities illegal unless comprehensive legal EU-US agreements are first in place.

Manual transfers

Traditionally, data have been manually transferred, and this is still an option for fast deployment. By manually transferring data, the investigator avoids the many pitfalls related to exposing possible sensitive data and research infrastructure to the public internet. However, plugging devices that have been in the hands of users, outside the confinements of the laboratory, into the research infrastructure to load data is also hazardous and must be done with great care. If the device has been infected with malware (e.g., viruses and ransomware) while in the hands of the user, the research infrastructure might suffer irreversible damage. Manually handling devices involves plugging in cables and copying data which is a time consuming and error prone process that requires significant human resources.

Faced with highly complex legislative issues in countries such as Norway regarding the regulation of personal data processing – commonly regarded as amongst the world's strictest – the technical possibilities associated with automated online data transfer can be restricted. Thus in the Norwegian arm of our study we opted for a traditional manual data transfer process via USB cables and portable storage media to on-premise institutional servers, but combined with smart devices purchased specifically for use in the project. Using devices owned by the research group has been a common practice, and in all but one of the twelve studies reviewed by Moore et al. (2017) reported that mobile devices were provided to the participants. We found that using devices owned by the research group, and thus not having participants use their own familiar devices can be very time-consuming in terms of device software initiation, data transfers, and the resetting of device software between participants. In particular, careful management is needed of the Apple ID-accounts required for downloading any iOS software, so as to not create a situation where participant information is leaked to unsuitable storage media related to personal accounts. Thus, even though we had produced a tool that was acceptable to users and appropriate to the purpose, manual data management on non-private devices via USB-devices to servers of the data controller using cable and USB-devices would be unsustainable on a long-term scale (i.e., over a period of years). Furthermore, adoption of such a manual approach renders it impossible to monitor data quality and provide feedback to participants in real-time, features that would be necessary for successful implementation in clinical settings. Even in projects with abundant human resources, manual transfer is an unattractive approach as human errors can easily compromise both data confidentiality and integrity.

Automatic transfers

Transferring data automatically over the internet is a tempting alternative to manual data transfer as much of the cumbersome process of physically handling devices is done by the participant. This was effectively executed in the US arm of our study. Upon completion of a testing session, the software application delivered detailed response data and audio files to a dedicated cloud service account, where performance measures were stored in a database. From this database, researchers could extract data, making it possible to effectively automate large amounts of data processing (e.g., for automatic speech recognition or semantic modeling of the language rich data), thus leveraging both resources in the cloud (e.g., Google's Speech-to-text API) or resources on the researchers own, on-premise hardware.

Establishing an online research infrastructure capable of handling data from participants' devices requires significant technical expertise and hardware investments. There are several emerging programming frameworks that can bring smart devices into the mainstream of psychological science (Piwek, Ellis, & Andrews, 2016), potentially avoiding large up-front investments by making use of available online cloud services (e.g., provided by Amazon, Microsoft, and Google). However, leveraging these resources are in many cases not possible due to the complicated legal and regulatory frameworks that govern all research on human participants. Regardless of whether data are collected as part of research or used in health services, it is expected that the responsible body has extensive knowledge of these binding legal frameworks. Indeed, the EU General Data Protection Regulation (European Union, 2016) adopted in May 2018 will be critically important for future ambulatory assessment tools within Europe as well as processing data collected from EU citizens, including processing by researchers or companies within for example the US. While the consequences of violating this regulation can be large (up to €20 Million or 4% of annual global turnover), the EU and respective national bodies provide a massive compilation of

guidelines, and in following these one can make large strides towards effective and safe systems. Notable sources of information include the European commission's site on the reform of the data protection rules in the EU (European Commission, 2018), and the handbook on security of personal data processing from the European Union Agency for Network and Information Security (2017). For mobile health application development in particular one can refer to the Code of Conduct on privacy for mobile health applications (European Commission, 2016). Within the time frame of many projects it can be difficult to establish all the numerous legal contracts required between research institutions, cloud providers and industry collaborators for a common international data management infrastructure allowing fully automated data transfers.

In some countries it is possible for research groups to purchase specialized infrastructure services that can enable them to quickly establish compliant data transfer, storage and processing. One such cloud service can serve as an illustration of necessary features for successful online implementations: In Norway, many universities subscribe to the Services for Sensitive Data at the University of Oslo. This allows researchers to store, view, and process their data by logging into a secure infrastructure using two-factor authentication. Each project is allocated its own virtual machine, a dedicated emulation of a suitable computer system hosted on servers running on university premises, connected to network storage system with secure backups. The service is designed to protect and ensure privacy of the respondents in compliance with EU laws and regulations.

Building on this established infrastructure, compliant mobile applications have been developed for research purposes within the health and social sciences, successfully transferring data from participants' own devices to secure storage in an automated fashion. Although many factors contribute to success, the following guidelines are key to overcoming regulatory challenges: (i) Using in-house development teams, rather than external third-party

software companies. This gives the project complete control over the source code and data flow. (ii) Sending data immediately (in an encrypted format) when the device is online. If offline, then encrypt and queue the data in temporary storage until the device is back online. (iii) Making data only available for analysis within a secure zone dedicated to the project. (iv) Using natural non-revealing descriptions of the application in app stores to avoid categorizing users (e.g., not listing an application as: "This is an application for patients with mental illness"). These fundamental features are necessary for confidentiality, integrity and regulatory compliance.

Swiftly achieving compliant cloud deployment of our software application by making use of such a services as the one provided by the University of Oslo would be possible, however, these models are often inherently one-way: Data can be sent from outside sources to the server, but communication from software running within the service cannot communicate back to the smart device. Security features like this simplifies the enforcement of confidentiality as participants only need their participant-ID to submit data, but in many research projects and health services there is an inherent need – and benefit – to provide timely feedback to the users based on data submitted. Future development and implementation of psychological assessment will be able to improve on usability design and privacy protection. This will be achieved by building on the foundation of the aforementioned services, and thereby ensure full control over the information collected, even when scientific progress pushes new data collection frameworks into legally uncharted terrain.

Legal issues

The legal issues that need to be considered when using a digitalized approach for longitudinal psychological assessment are linked to the right to privacy, unequivocally established in Article 12 of the Human Rights Declaration (UN General Assembly, 1948), and

can be regarded as two-folded: First, since the data collection and processing will span periods where opinions may change, participants may wish to withdraw their consent for participation and thus retract their data. This may be especially an issue in patients with serious mental illness whose mental states may by the very nature of their illness fluctuate. Second, the very nature of the data collected and analysis performed may be opaque to participants, challenging the notion that consents are conducted with true knowledge of the scope of the contract that is agreed upon. We discuss below how designing and implementing rigorous privacy policies and data de-classification pathways can help to improve the technical availability of data so as to make it possible to move the psychological assessment out of the lab and into the hands of the individual, while complying with the strictest legal standards.

Managing informed consent

Recent regulation trends point towards a strengthening of individuals' right to control over information regarding themselves, including the right to have their private data deleted, often expressed as the "right to be forgotten" (European Union, 2014, 2016). For medical research, these new rights represent a difficult challenge because of the sheer volume of data collected, and the complex data sharing patterns and plethora of data processing tools often necessary. Indeed, it is not uncommon for a researcher to have data entered in multiple spreadsheets, copied and distributed between researchers and devices. Therefore, it quickly becomes a practically impossible task to track down and delete all entries from a single individual who revokes their consent. Many centralized data repositories now ensure that sensitive data are not copied outside the secure infrastructure. Still, they must all allow for the natural flow of data between researchers and their tools, and provide no holistic means to track or control data within the boundaries of their systems.

Solutions to this lie in the design of the consent process, data collection tool, storage system, and analysis tools. The common practice of using one-off, paper-based consent forms with fixed statements on the purpose of research data has come under scrutiny, and the concept of "dynamic consent" has been proposed as a solution (Kaye et al., 2012, Kaye et al., 2015; Budin-Ljøsne et al., 2015, 2017). This concept proposes interactive personalized interfaces where individuals can engage with research groups and alter their consent choices in real time (Kaye et al., 2015). While awareness around these issues has mostly come from biobank research, where broad and long-term consents are common, this is highly relevant in future implementations of longitudinal psychological assessment frameworks. With more clearly defined consents, data management can be a technical process of applying privacy policies closely connected to what is collected throughout the life-cycle of the data. Privacy policies are rules that define what can be done with the data, for example who can have access to read or change a file, and for what purpose. Newly proposed mechanisms enable users to define and attach highly customized privacy policies as metadata (Johansen et al., 2015). Such mechanisms must support policies that can change depending on how data are manipulated, apply policies to all copies of data and to any derived data. To ensure compliance, the underlying computer infrastructure must enforce such policies at the system level.

Tracking and controlling the information flow within a computer system is a mature topic in computer science, and fine-grained control of information flow is possible by instrumenting the source code with policy labels (Sabelfeld & Myers, 2003). However, this does not work for the many existing analytical tools that researchers employ today. Controlling information flow at the operating system is a more realistic approach for a research infrastructure as applications do not need to be rewritten. Although several academic systems have been demonstrated (Efstathopoulos et al., 2005; Enck et al., 2010), an off-the-shelf solution suitable for research has yet to emerge. A more practical approach is to attach

policy labels to files (e.g., Johansen et al., 2015). With this, each policy label identifies a state in a per-user privacy robot (a so-called Privaton) that grants or denies access based on the stated purpose of processing. The system can then check that the purpose of processing match the one granted by users' consent. Such privacy labels can be attached to data when created and made inseparable from that data, even when uploaded to a remote storage infrastructure.

While existing institutional infrastructure can be well-protected using traditional security mechanisms like encryption, firewalls, and multi-factor authentication, systems designed for true personal control over one's own data will likely be a core aspect of legally moving psychological assessment out of the lab and into the hands of the individual. By enabling this individual control, a technology that can potentially be experienced as invasive to privacy can actually instead result in personal empowerment.

Escalating sensitivity of data

The sheer volume and unique possibilities of combining data means that there is the possibility that previously trivial data can suddenly turn into highly sensitive information. This can be an additional complication to the consent process, as it can be difficult to predict the level of sensitivity of detailed and voluminous longitudinal data. For example, results from the Stroop Color Word task would probably be considered to provide fairly 'mundane' information regarding an individual's attentional abilities. However, the sheer detail that now can be collected via a smart device means that this information can be translated into highly accurate timing information which could, for example in certain clinical scenarios be indicative of extreme anxiety or the onset of mania, which thus additionally puts the onus and burden on the investigator to ensure timely feedback rather than analyze the data several months later. Consider a hypothetical scenario where a participant's data from our mobile application reveals erratic touch and timing responses on several tasks, and additionally has

high pitch ‘anxious-sounding’ voice recordings from the Stroop test, and possibly extreme values on other task measures such as from self-reported negative affect. These are all important clinically and are in line with the health goals of the research. However, *in combination* they clearly are strongly indicative of the necessity of an immediate action on part of the investigator, and a timely response in this case would be to activate an emergency response by contacting the patient and relevant health professionals for increased patient care. The very combination of singular metrics has now escalated the dataset to a level of potentially highly sensitive - and valuable - health information. Furthermore, the fact that the device can record speech means that the participant’s speech during a speaking task such as the Stroop task - no matter what they say (e.g., ‘I feel extremely depressed and this task is very irritating and I want to die’) - will also be recorded. This data has now suddenly escalated in terms of sensitivity even though certain information was not even solicited. This hypothetical example serves to further emphasize the need for a dynamic system for managing data and consent issues.

In the context of the specific case of the software application we developed, where vocal responses were central, collecting speech for acoustic and semantic analyses introduced another complex and specific privacy-related challenge: High-quality speech recordings can in themselves lead to direct identification of an individual, and in addition, due to the ambulatory setting there are no easy ways to ensure that other identifying information will not end up in the resulting dataset. An important procedure for declassifying datasets is removing any links to directly identifiable information such as names and contact information. However, the inherent characteristics of speech data, combined with the richness from the multiple data streams that can be collected with mobile psychological assessment tools, makes the risk of re-identification of participants high. As methods of analysis become more developed, and the processing power on smart devices increases, it will be feasible to extract

more data from responses before transferring to the institution. Consider our Stroop-example, where the speech recognition that currently relies on cloud services and on premise hardware may in the near-future be analyzed and time-stamped on the device itself, making the need to transfer identifiable speech samples obsolete.

Understanding how to classify the data derived from mobile applications, specifically to what extent it would qualify as 'health information', will be an important first step in establishing the legal and technical frameworks necessary for any implementation. Singular measures were considered to be analogous with innocuous gaming scores, but multivariate results could - and should - eventually form the basis for a description of 'health and illness'. By having the intent to produce health information, data collection exceeds a threshold and enters the strictest legal domain from the very onset of such study design. Even when the level of sensitivity is unknown, as can be common in many research settings, defaulting to a higher classification is wise and proactive.

For many researchers, the challenge with this escalation of sensitivity when conducting remote and daily assessment will be two-fold. First, limited technical and computing expertise renders several approaches as simply not feasible due to technical constraints. Second, higher levels of expertise in both psychological assessment and technical implementation make it possible to conduct innovative research, but can result in situations where legal frameworks are strained, in particular when it comes to informed consent in vulnerable persons (e.g., in some clinical populations). This combination of innovation in psychological-, technical- and legal sciences provide interesting venues for progress in the years to come.

Cultural issues

Successful implementations of technology to enable remote and frequent psychological assessment can be scalable and provide the foundation for new insights into behavior and how it is affected by illness. However, such a bold approach requires careful consideration of languages, cultures and other individual difference factors. In our software applications, we had a specific focus on assessing and parsing specific aspects of cognition via the medium of language. Such a deconstructivist approach to cognition - and its dysfunction - is not novel. However, to deconstruct the underlying processes in language - which themselves may differ across languages - is arguably challenging but likely to yield extremely rich data of importance to behavioral science. Language is deeply affected by culture, but what is less clear is exactly how cognition, often expressed and interpreted through language, may be differentially modulated by cultures. Obviously tasks need to be suitably translated (and back-translated) and normed within the various languages and cultures that the tasks are to be implemented. However, beyond these relatively obvious task design issues it is also necessary to establish that the resulting tasks fit well given cultural variations, both in terms of what we expect to observe regarding use of language as well as how contextual factors may affect verbal behavior differently across cultures. Clearly there is a risk that the putative differences can seem large, but within an assessment design that focuses on relative change within individuals in a longitudinal framework, many of the expected differences between national versions of tests may not completely negate the value of such tools. Nonetheless, in designing these tools to assay psychological functions we must expend great effort to ensure the tasks are language-neutral and culture-fair.

Language-specific issues

With the shared traditions of US and European psychological assessment methods, translating task instructions and content from our English version of the software application to a Norwegian version was mostly straightforward. However, an example from translating the Stroop test serves to illustrate how cross-cultural implementations can affect core psychophysical properties of the very behavioral effects under investigation. In the single-trial paradigm we adopted (based on Perlstein et al. 1998), the English stimuli included color-words of 3-, 4, -5 and 6-letter lengths counterbalanced across sessions (i.e., RED, BLUE, GREEN and PURPLE respectively). Directly translating these colors from English would have yielded Norwegian words of 3, 3, 5 and 5 letter lengths (i.e., RØD, BLÅ, GRØNN and LILLA) and thus not be comparable in terms of similar word lengths. However, finding a commonly used 6-letter color-word in Norwegian proved challenging, and we selected the low-frequency word TURKIS (turquoise) in the Norwegian version (Figure 2 Panel B). A caveat with this color is that visually it can be perceived as “GREEN” or “BLUE” by the participants (Figure 2 Panel C). In such a case, the participant presented with the stimulus TURKIS may experience the ink-color as incongruous, thus introducing a possible unintentional interference effect in these stimuli. Changing stimulus characteristics in the Norwegian version to avoid this issue, specifically such that the word lengths would not be comparable, was considered a more problematic methodological modification, and thus the turquoise color was implemented. Beyond this practical, methodological issue, it is worth noting that even if the Stroop task is not very complicated in terms of the actual language used, cross-cultural differences have been reported (e.g., Magiste, 1985) and certain languages show a difference in the magnitude of the interference effect compared to others (Alnasari & Baroun, 2004).

Computational language methods afford an approach to psychological assessment in language that goes beyond simple word counts. For example, in the widely used category fluency task (e.g., 'Name as many ANIMALS as you can in a minute'), it is possible to actually assess the flow and meaning of an utterance (Elvevåg et al., 2007; Nicodemus et al, 2014), and to measure differences across languages that may reveal important clinical markers potentially missed without careful cultural consideration (e.g., in Norway it might be more usual to list the names of many fish given many peoples' geographical proximity to the coast, whereas this might be considered unusual in many cultures and languages who are either not near the coast or more widely distributed geographically). Our previous experience of building analytical (semantic) tools in the Norwegian language revealed how dominant English words can be such that they penetrate into other languages (Rosenstein et al., 2015). A notable example is the global use of the English word 'and' (e.g., 'rock *and* roll') but in the case of Norwegian 'and' translates into the bird 'duck'. So dominant is the use of this word within English phrases that are thus adopted also within non-English languages, that even in cases where it might mean something very different it can introduce unexpected error in behavioral and cognitive models and so must be addressed. In the case of our previous semantic modeling of animal fluency words in Norwegian we caught such instances by using the text categorization technique of Cavnar and Trenkle (1994) on small windows around "and" to separate English "rock and roll" occurrences from Norwegian "Sprø and med appelsin og koriander" (Translation: "crispy duck with orange and coriander")" (p.127; Rosenstein et al. 2015).

Beyond the aforementioned methodological design issues in our Stroop example, this task also illustrates how the actual analysis - and in our case the automatization of the task - can be differentially affected. In our project we sought to use Automatic Speech Recognition (henceforth ASR) to fully automate the task and its analysis (Holmlund et al, 2018). The value

for psychological assessment of embracing ASR is that the person's speech can be captured by the smart device's microphone, converted to a digital signal, and then recognized by an ASR system which then can produce a sequence of the words along with ancillary information (e.g., the timing of the words and other speech-related events such as pauses and disfluencies). This time-aligned ASR will give the most likely sequence of words based on the sophistication of the actual language models used. In the English version of our Stroop task, we leveraged the enormous benefit afforded by Google's English speech model as well as using a language model specifically tuned to recognize the relevant words in the Stroop task specifically (i.e., the color words) such that the word error rate was approximately 6%, which is fairly accurate. Use of ASR could revolutionize the manner in which such core cognitive processes are assessed in both research and clinical settings as well as challenge existing cognitive neuroscientific models that currently exist (Holmlund et al, 2018). However, the prognosis for such fully automated tools in non-English languages that leverage ASR tools is less clear and likely requires many years and much effort of collecting text corpora to first build the appropriate language analysis tools. This advantage of the English language, courtesy of its dominance, is parallel to the phenomenon where psychological assessment tools developed for the English language can reach maturity and sophistication more quickly.

Collating data across countries: Future possibilities

The increasing use of personal digital assistants that require speech interaction increases the need for having computational models of language, something that can be leveraged in cognitive and behavioral research. With a demand for speech recognition, machine translation or semantic models for a language, large industry forces have an incentive to develop multilingual methods. An example of this is the MUSE-library (Conneau, Lample, Ranzato, Denoyer & Jégou, 2017) recently published by the social media

company Facebook and relevant to language research that employs vector-space models, such as our own work in the semantic analysis of story retellings (Foltz et al., 2018). While this technology has been developed mainly for machine translation, the methods can create word vectors for different languages (e.g., English and Norwegian) aligned within the same vector space, possibly providing a common ground for analyses. Advances in these methods will probably first be seen for big languages such as English, and then subsequently for the smaller languages, although naturally this will be faster in countries who devote resources to such developments. Additionally, these methods will benefit from the fact that it will become easier to collect language data as more and more communication and knowledge repositories enter the online realm.

Conclusions

New technological frameworks provide unprecedented opportunities for remote self-administered behavioral and clinical assessments, where it is possible to participate in easy-to-use digital versions of traditional behavioral tests as well as new variants that are suitable for use on a daily basis. However, employing such a methodological approach, both locally and internationally, necessitates that the technological infrastructure is sufficiently secure so as to ensure the safety and integrity of data transfers. Manually moving data between hardware devices is labor intensive, and although moving data via internet infrastructures is much more efficient, it demands adherence to the strict legal frameworks that regulate such transfers within and across the countries involved. These same legal frameworks also grant participants strong rights to their own data, and they can request deletion of their data at any point, and thus this necessitates development of quite a sophisticated data management infrastructure. Design of assessments must consider the usability of the items across countries and cultures as well as how language may influence performance. Sensitivity to language is not just a

matter of accurate translation of assessment items, but must also incorporate a deep understanding of how participants use the language and how that may affect analysis and interpretation of results since many assessments rely on language for comprehension and expression. In sum, to fully harness the power of this new technological approach, research needs to be increasingly multidisciplinary - methodologically, technically, legally and culture-sensitive - so as to ensure high levels of user satisfaction and superior data quality and to guarantee the highest possible level of protection of each participant's resulting dataset. The scientific and clinical value of successfully moving psychological assessment out of the controlled laboratory setting affords an unprecedented opportunity to explore the temporal dynamics underlying human behavior and to understand more completely individual differences given the multiple channels of behavior that can be simultaneously sampled.

References

Alansari, B., & Baroun, K. (2004). Gender and cultural performance differences on the Stroop Color and Word Test: A comparative study. *Social Behavior and Personality, 32*(3), 235-245. doi: 10.2224/sbp.2004.32.3.233

Allard, M., Husky, M., Catheline, G., Pelletier, A., Dilharreguy, B., Amieva, H., . . . Swendsen, J. (2014). Mobile technologies in the early detection of cognitive decline. *PLoS ONE, 9*(12), 1-10. doi:10.1371/journal.pone.0112197

Anderson, K., Burford, O., & Emmerton, L. (2016). Mobile health apps to facilitate self-care: A qualitative study of user experiences. *PLoS ONE*, *11*(5), 1-21. doi:10.1371/journal.pone.0156164

Au, R., Piers, R., Devine, S., & Brown, G. (2017). How technology is reshaping cognitive assessment: Lessons from the Framingham heart study. *Neuropsychology, 31*(8), 846-861. doi: 10.1037/neu0000411

Bauer, R., Iverson, G., Cernich, A., Binder, L., Ruff, R., & Naugle, R. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology, 27*(3), 362-373. doi: 10.1093/arclin/acs027

Ben-Porath, Y. (2016). Inaugural editorial for Psychological Assessment [Editorial], *Psychological Assessment, 28*(1), 1-2. doi: 10.1037/pas0000270

Bilder, R. (2011). Neuropsychology 3.0: Evidence-based science and practice. *Journal of the International Neuropsychological Society, 17*(1), 7-13. doi: 10.1017/S1355617710001396

Brouillette R. M., Foil H., Fontenot S., Correro A., Allen R., Martin C. K., … Keller, J. N. (2013) Feasibility, reliability, and validity of a smartphone based application for the

assessment of cognitive function in the elderly. *PLoS ONE, 8*(6), 1-5. doi:10.1371/journal.pone.0065925

Budin-Ljøsne, I., Bentzen, H. B., Solbakk, J. H., & Myklebost, O. (2015). Genome sequencing in research requires a new approach to consent. *Tidsskriftet for den Norske Laegeforening, 135*(22), 2031-2032. doi:10.4045/tidsskr.15.0944

Budin-Ljøsne, I., Teare, H., Kaye, J., Beck, S., Bentzen, H., Caenazzo, L., . . . Mascalzoni, D. (2017). Dynamic consent: A potential solution to some of the challenges of modern biomedical research. *BMC Medical Ethics, 18*(1), 4. doi:10.1186/s12910-016-0162-9

Cavnar, W. B. & Trenkle, J. M. (1994). N-gram-based text categorization. Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 161–175. Retrieved from http://odur.let.rug.nl/vannoord/TextCat/textcat.pdf

Cohen, A. S., Schwartz, E. K., Mitchell, K. R, Le, T., Foltz, P. W., Holmlund, T. B., & Elvevåg, B. (2018a). *21st Century psychometrics for circumventing the "Psychiatric Miasma".* Manuscript submitted for publication.

Cohen, A. S., Fedechko, T., Schwartz, E., Le, T., Foltz, P. W., Bernstein, J., . . . Elvevåg, B. (2018b). *Psychiatric risk assessment from the clinician's perspective: Lessons for the future.* Manuscript submitted for publication.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). *Word translation without parallel data*. Retrieved from Arxiv database. (ID: 1710.04087)

Datatilsynet. (2017). Personal Data Act, Act of 14 April 2000 No. 31 relating to the processing of personal data. Retrieved from https://www.datatilsynet.no/en/regulations-and-tools/regulations-and-decisions/norwegian-privacy-law/personal-data-act/

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. -X., . . . Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can

revolutionize research in cognitive science. *PLoS ONE, 6*(9), 1-3. doi:10.1371/journal.pone.0024974

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment, 21*(4), 463. doi: 10.1037/a0017075

European Commission. (2016). Code of Conduct on privacy for mHealth apps has been finalised. Retrieved from https://ec.europa.eu/digital-single-market/en/news/code-conduct-privacy-mhealth-apps-has-been-finalised

European Commission. (2018). 2018 reform of EU data protection rules. Retrieved from https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

European Union. (2014). Factsheet on the "Right to be forgotten" ruling (C-131/12). Retrieved from http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (Official Journal of the European Union, Vol. L119, pp. 1-88). Retrieved from http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679

European Union. (2017). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. (Official Journal of the European

Union, Vol. L117, pp. 1-175). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0745

European Union Agency for Network Information Security. (2017). Handbook on security of personal data processing. Heraklion: ENISA. doi: 10.2824/569768

Efstathopoulos, P., Krohn, M., VanDeBogart, S., Frey, C., Ziegler, D., Kohler, E., . . . Morris, R. (2005). Labels and event processes in the asbestos operating system. *SIGOPS Operating Systems Review, 39*(5), 17-30. doi:10.1145/1095809.1095813

Elvevåg, B., Foltz, P., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia Research, 93*, 304-316. doi:10.1016/j.schres.2007.03.001

Enck, W., Gilbert, P., Chun, B.-G., Cox, L. P., Jung, J., McDaniel, P., & Sheth, A. N. (2010). *TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones*. Proceedings of the 9th USENIX conference on Operating systems design and implementation, Vancouver, BC, Canada. Retrieved from https://www.usenix.org/legacy/event/osdi10/tech/full_papers/Enck.pdf

Foltz, P. W., Cheng, J., Bernstein, J. C., Rosenfeld, E. P., Cohen, A. S., Holmlund, T. B., & Elvevåg, B. (2018) *Natural language processing for automating the assessment of verbal memory: Challenges and opportunities*. Manuscript submitted for publication.

Frings, L., Wagner, K., Carius, A., Schinkel, A., Lehmann, C., Schulze-Bonhage, A., & Maiwald, T. (2008). Early detection of behavioral side effects of antiepileptic treatment using handheld computers. *Epilepsy and Behavior, 13*(2), 402-406. doi:10.1016/j.yebeh.2008.04.022

Golden, C. J. (1976). Identification of Brain Disorders by Stroop Color and Word Test. *Journal of Clinical Psychology, 32*(3), 654-658. doi:10.1002/1097-4679(197607)32:3<654::Aid-Jclp2270320336>3.0.Co;2-Z

Huckvale, K., Prieto, J., Tilney, M., Benghozi, P., & Car, J. (2015). Unaddressed privacy risks in accredited health and wellness apps: A cross-sectional systematic assessment. *BMC Medicine, 13*(1), 214, 1-13. doi:10.1186/s12916-015-0444-y

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S., Bernstein, J., Rosenfeld, E., . . . Elvevåg, B. (2018). *Deconstructing attentional control: Automated analysis of a spoken Stroop task.* Manuscript submitted for publication.

Johansen, H. D., Birrell, E., Renesse, R. v., Schneider, F. B., Stenhaug, M., & Johansen, D. (2015). *Enforcing privacy policies with meta-code*. Proceedings of the 6th Asia-Pacific Workshop on Systems, Tokyo, Japan. doi: 10.1145/2797022.2797040

Jongstra, S., Wijsman, L. W., Cachucho, R., Hoevenaar-Blom, M. P., Mooijaart, S. P., & Richard, E. (2017). Cognitive testing in people at increased risk of dementia using a smartphone app: The iVitality proof-of-principle study. *JMIR mHealth and uHealth*, 5(5), 1-11. doi:10.2196/mhealth.6939

Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S., Kanellopoulou, N., . . . Winter, S. (2012). From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews: Genetics, 13*(5), 371-376. doi:10.1038/nrg3218

Kaye, J. A., Morrison, M., Teare, H., Melham, K., Whitley, E., & Lund, D. (2015). Dynamic consent: A patient interface for twenty-first century research networks. *European Journal of Human Genetics, 23*(2), 141-146. doi:10.1038/ejhg.2014.71

Kennedy, D. O., Veasey, R. C., Watson, A. W., Dodd, F. L., Jones, E. K., Tiplady, B., & Haskell, C. F. (2011). Vitamins and psychological functioning: A mobile phone assessment of the effects of a B vitamin complex, vitamin C and minerals on cognitive performance and subjective mood and energy. *Human Psychopharmacology, 26*(4–5), 338–347. doi:10.1002/hup.1216

Kindt, M., Bierman, D., & Brosschot, J. F. (1996). Stroop versus stroop: Comparison of a card format and a single-trial format of the standard color-word stroop task and the emotional stroop task. *Personality and Individual Differences, 21*(5), 653-661. doi:10.1016/0191-8869(96)00133-X

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin, 109*(2), 163-203. doi:10.1037/0033-2909.109.2.163

Mägiste, E. (1985). Development of intra- and interlingual interference in bilinguals. *Journal of Psycholinguistic Research*, *14*(2), 137-154. doi:10.1007/BF01067626

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

Neurological Devices, 21 C.F.R. §882.1470. (2017). Retrieved from https://www.ecfr.gov/cgi-bin/text-idx?SID=a022c48692499a5cc054f60f223f83b0&mc=true&node=se21.8.882_11470&rgn=div8

Nicodemus, K.K., Elvevåg, B., Foltz, P.W., Rosenstein, M., Diaz-Asper, C., & Weinberger, D.R. (2014). Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex*, *55*, 182-191. doi:10.1016/j.cortex.2013.12.004

Nielsen, J. (1993). *Usability engineering*. Boston, Mass: Academic Press.

Perlstein, W. M., Carter, C. S., Barch, D. M., & Baird, J. W. (1998). The stroop task and attention deficits in schizophrenia: A critical evaluation of card and single-trial stroop methodologies. *Neuropsychology, 12*(3), 414-425. doi:10.1037//0894-4105.12.3.414

Piwek, L., Ellis, D. A., & Andrews, S. (2016). Can programming frameworks bring smartphones into the mainstream of psychological science? *Frontiers in Psychology, 7*, 1-6. doi:10.3389/fpsyg.2016.01252

Rabin, L., Spadaccini, A., Brodale, D., Grant, K., Elbulok-Charcape, M., Barr, W., & Brown, Ronald T. (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Professional Psychology: Research and Practice, 45*(5), 368-377. doi: 10.1037/a0037987

Riediger, M. G., Wrzus, C., Klipker, K., Müller, V., Wagner, G., & Schmiedek, F. (2014). Outside of the laboratory: Associations of working-memory performance with psychological and physiological arousal vary with age. *Psychology and Aging, 29*(1), 103-114. doi:10.1037/a0035766

Rosenstein, M., Foltz, P. W., Vaskinn, A., & Elvevåg, B. (2015). Practical issues in developing semantic frameworks for the analysis of verbal fluency data: A Norwegian data case study. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 124–133. Denver, Colorado, June 5, 2015. Association for Computational Linguistics. Retrieved from http://m-mitchell.com/clpsych2015/pdf/CLPsych15.pdf

Sabelfeld, A., & Myers, A. C. (2003). Language-based information-flow security. *IEEE Journal on Selected Areas in Communications, 21*(1), 5-19. doi:10.1109/Jsac.2002.806121

Schuster, R., Mermelstein, R., & Hedeker, D. (2015). Acceptability and feasibility of a visual working memory task in an ecological momentary assessment paradigm. *Psychological Assessment, 27*(4), 1463. doi:10.1037/pas0000138

Schweitzer, P., Husky, M., Allard, M., Amieva, H., Peres, K., Foubert-Samier, A., . . . Swendsen, J. (2017). Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. *International Journal of Methods in Psychiatric Research (Online), 26*(3), 1-8. doi:10.1002/mpr.1521

Shiffman, S. (2009). How many cigarettes did you smoke? Assessing cigarette consumption by global report, Time-Line Follow-Back, and ecological momentary assessment. *Health Psychology, 28*(5), 519. doi: 10.1037/a0015197

Sliwinski, M. J., Mogle, J. A., Hyun, J. M., Munoz, E. B., Smyth, J., & Lipton, R. (2016). Reliability and validity of ambulatory cognitive assessments. *Assessment,* 25(1), 14-30. doi:10.1177/1073191116643164

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662. doi:10.1037//0096-3445.121.1.15

Tieges, Z., Stíobhairt, A., Scott, K., Suchorab, K., Weir, A., Parks, S., . . . MacLullich, A. (2015). Development of a smartphone application for the objective detection of attentional deficits in delirium. *International Psychogeriatrics, 27*(8), 1251-1262. doi:10.1017/S1041610215000186

Timmers, C., Maeghs, A., Vestjens, M., Bonnemayer, C., Hamers, H., & Blokland, A. (2014). Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology: Adult*, *21*(2), 136-142. doi:10.1080/09084282.2013.778261

Tiplady, B., Oshinowo, B., Thomson, J., & Drummond, G. (2009). Alcohol and cognitive function: Assessment in everyday life and laboratory settings using mobile phones. *Alcoholism - Clinical and Experimental Research*, *33*(12), 2094-2102. doi:10.1111/j.1530-0277.2009.01049.x

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

UN General Assembly. (1948). "Universal declaration of human rights" (217 [III] A). Paris. Retrieved from http://www.un.org/en/universal-declaration-human-rights/

United States. (2004). The Health Insurance Portability and Accountability Act (HIPAA). Washington, D.C.: U.S. Dept. of Labor, Employee Benefits Security Administration.

Waters, A., & Li, Y. (2008). Evaluating the utility of administering a reaction time task in an ecological momentary assessment study. *Psychopharmacology*, *197*(1), 25-35. doi:10.1007/s00213-007-1006-6

Westerhausen, R., Kompus, K., & Hugdahl, K. (2011). Impaired cognitive inhibition in schizophrenia: a meta-analysis of the Stroop interference effect. *Schizophrenia Research*, *133*(1-3), 172-181. doi:10.1016/j.schres.2011.08.025

Witt, J. -A., Alpherts, W., & Helmstaedter, C. (2013). Computerized neuropsychological testing in epilepsy: Overview of available tools. *Seizure: European Journal of Epilepsy*, *22*(6), 416-423. doi: 10.1016/j.seizure.2013.04.004

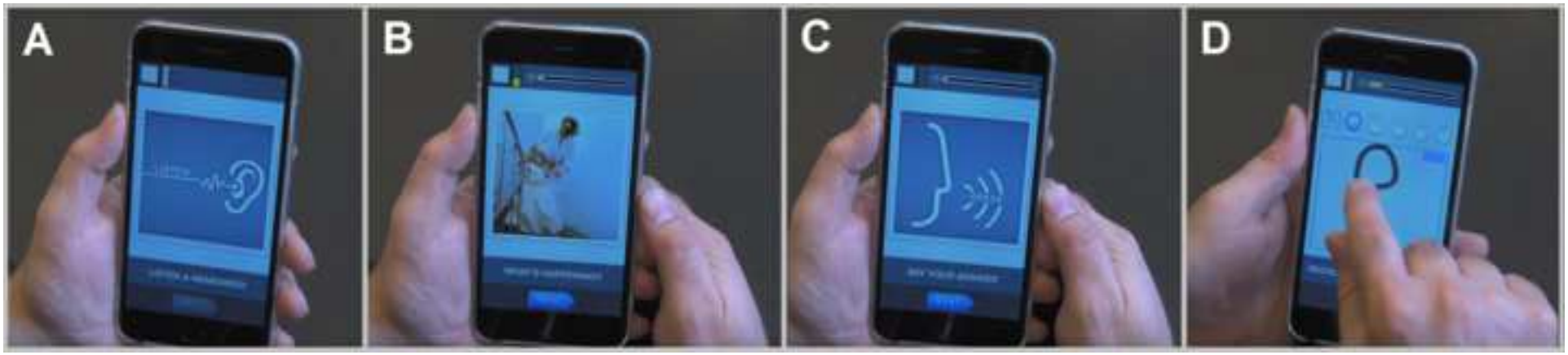
Yang, C. H., Maher, J. P., & Conroy, D. E. (2015). Acceptability of mobile health interventions to reduce inactivity-related health risk in central Pennsylvania adults. *Preventive Medicine Reports*, *2*, 669-672. doi:10.1016/j.pmedr.2015.08.009

Figure legends

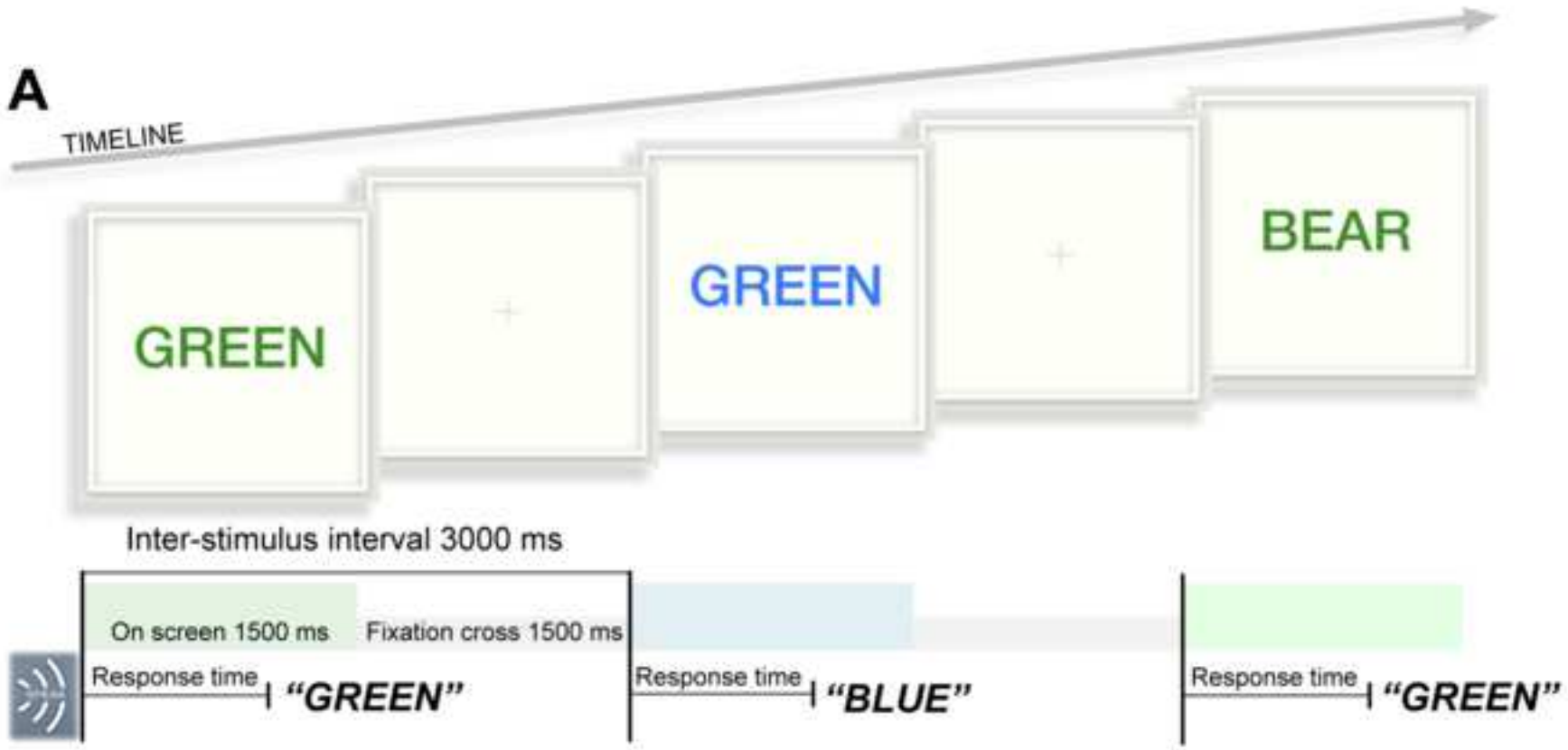
Figure 1. The tasks were short and engaging and required listening (Panel A), watching (Panel B), speaking (Panel C), and touching (Panel D) to interact with the smart device.

Figure 2. Panel A: Thirty-two words of four possible colors were presented every 3000 ms and stayed on the screen for 1500 ms. The task was to say out loud, as fast as possible, the color of the word, ignoring what was written, and thus the correct answers above would be GREEN, BLUE and GREEN respectively. Vocal responses were recorded by the microphone on the device, and saved in a file for timestamping using automatic speech recognition software. Panel B: Adapting the task to Norwegian language made it necessary to change the actual colors in order to maintain stimulus specifications, such as the number of characters in words, but this introduced some trade-offs to ensure consistency between versions. (Translation of Norwegian stimuli: rød = red, brun = brown, lilla = purple, turkis = turquise, ape = monkey, hund = dog, tiger = tiger, slange = snake). Panel C: An illustration of a challenge with some color-choices may be sub-optimal visibility on some screens and in some lighting conditions. This screenshot presents the word PURPLE ('lilla' in Norwegian) in the color turquoise that was employed in the Norwegian version ('turkis' in Norwegian).

Masked Figure 1



Masked Figure 2



B

Stimuli in US version

RED
BLUE
GREEN
PURPLE
DOG
BEAR
TIGER
MONKEY

Stimuli in Norwegian version

RØD
BRUN
LILLA
TURKIS
APE
HUND
TIGER
SLANGE

C