UiT The Arctic University of Norway

Faculty of Health Sciences

Department of Clinical Medicine

Modeling remotely collected speech data

Applications for psychiatry

Terje Bektesevic Holmlund

A dissertation of the degree of Philosophiae Doctor, October 2019

Table of Contents

1 Foreword ... 3

2 Abbreviations ... 5

3 Abstract ... 6

4 Sammenfatning på norsk (Norwegian summary) ... 7

5 List of papers ... 9

6 Introduction ... 10

6.1 Speech and language in psychiatry ... 13

6.2 Neuropsychological assessment using speech responses ... 15

6.2.1 The Stroop Color Word test and measurements of attentional bias and control. ... 16

6.2.2 Verbal fluency tests can provide measurements of the flow and timing of multiple spoken words. ... 17

6.2.3 Retelling a story can enable measurements of verbal memory ability ... 19

6.3 Two core analytical methods for analysis of speech production and semantic content ... 20

6.3.1 Automatic Speech Recognition for assessing speech production ... 21

6.3.2 Natural Language Processing and word vectors for semantic analysis ... 22

6.4 Introducing technology in psychiatry ... 26

6.5 Natural language processing for research and clinical utility in psychiatry ... 28

6.6 The aims of the thesis ... 29

7 Methods ... 32

7.1 The *d*MSE and the *MinTest* mobile applications ... 32

7.2 Participants ... 33

7.3 Procedure and analysis ... 35

7.3.1 Paper I: Discussion of anecdotes from practical implementation ... 35

7.3.2 Paper II: Temporal measurements of single-word speech production in the Stroop color-word task ... 36

7.3.3 Paper III: Timestamps and semantic vectors for multiple word sequences in a verbal fluency test ... 39

7.3.4 Paper IV: Verbal memory recall of stories ... 40

8 Ethics ... 43

9 Results ... 45

9.1 Paper I: Moving assessment out of the lab - Infrastructure for data capture ... 45

9.2 Paper II: Single word speech production in the Stroop task ... 50

9.3 Paper III: Speech production and semantic coherence in the verbal fluency task 52
9.4 Paper IV: Computational analysis of recall performance in the verbal memory task 56
10 Discussion 58
10.1 Limitations 60
10.2 What will these new technological approaches mean for psychiatry? 64
11 Conclusion 68
12 Final remarks and future perspectives 68
13 References 72
Papers I-IV
Paper I: Moving psychological assessment out of the controlled laboratory setting: Practical challenges
Paper II: Using automated speech processing for repeated measurements of attentional bias and control
Paper III: Updating verbal fluency analysis for the 21st century: Applications for psychiatry
Paper IV: Applying speech technologies to assess verbal memory in patients with serious mental illness
Appendix 1: Written instructions for the *d*MSE
Appendix 2: Online questionnaire for *MinTest* participants
Appendix 3: Consent form - *d*MSE
Appendix 4: Consent form - *MinTest*

List of Tables

Table 1 - Participants 35

List of Figures

Figure 1. 25
Figure 2. 33
Figure 3. 37
Figure 4. 43
Figure 5. 47
Figure 6. 49
Figure 7. 50
Figure 8. 51
Figure 9. 55
Figure 10. 69

1 Foreword

Words are important. Even a single spoken word can have a dramatic effect on our surroundings: "*Stop!*". Combinations of words can convey complex ideas, plans, intent, feelings, affecting others: "*You are doing great, keep moving forward!*". Words are behavioral atoms whose conceptual combination affords an almost infinite number of ideas. However, even the combinations need context, a space where they can act. While written words can be effective for those distant in both place and time, spoken words are for immediate effect, be that for shouting a message across a busy street or for a quiet conversation with a loved one. In the case of my research, the conversation may be between a patient and a clinical caretaker. The spoken words are ephemeral, appearing and then dissipating in our recollection of events. Meanings get blurred and lost with time. This is the way it should be, most of the time. At other times more precision is needed. Resolution, robustness, accountability and objective analysis is required in the realms of medicine and science. This is when we need engineering.

Digital representations are useful. Transforming the physical sound pressure waves of speech into codes of zeros and ones make quantitative analysis possible. We can do computations. The numbers allow us to grab the word, inspect it, and each millisecond of an utterance can be available for detailed examination. A word frozen in time. Words and sentences spoken several years ago on another continent, North America, are the object of study in this investigation, available for dissection by numerical process. In the following thesis I will argue that we have enough tools available to do deep mapping of mental states using analysis of spoken words. This is research into measurements of how we speak.

The research would not have been possible without the large group of wonderful and brilliant people. First of all I would like to thank my supervisor Brita Elvevåg. It has been a great honour to be able to work with one of the founders of the field of computational language analysis in psychiatry. In addition, she is immensely kind and patient and has given me more hours of mentoring than any student can hope for. Thanks also to my co-supervisor Bruno Laeng. His steady mentorship has kept me inspired over many years, and his vast knowledge of cognitive processes and experimental procedures has been hugely important in shaping my understanding of cognitive neuroscience.

I am also tremendously grateful for the welcome I received in the research group that created the main instrument of this thesis; the *delta* Mental State Examination. Among these are the esteemed scientists Alex S. Cohen, Peter W. Foltz, Jared Bernstein, Jian Cheng and Elizabeth Rosenfeld. Additionally I was lucky to receive the support of Håvard Johansen and Randi Sigurdsen at the Arctic University of Norway, as well as Dagfinn Bergsager and Pål Fugelli at the Services for Sensitive Data at the University of Oslo.

On a daily basis I received crucial support from the wonderful people of the corridor at the Åsgård psychiatric hospital. Thank you for inspiring talks and delicious Friday lunches. A particularly honourable mention is deserved to Joaquim Carvalho for keeping everything in perfect order around us, a true master of infrastructure, an enemy of entropy.

Several other students and friends must be mentioned. Chelsea Chandler, Taylor Fedechko, Thanh P. Le and Tovah Cowan have graciously included me in the production of their magnificent papers and conference contributions. I want to thank Connie Malen Moen for her contributions in the early stages of development of the Norwegian *MinTest* mobile application. A huge thanks to all the participants hwo supplied us with all the wonderful data: we will make good use of it!

For my family and friends: Thank you for your continuous support. Thanks to my parents who made me who I am, a congenital scientist. To my beautiful, intelligent and particularly awesome wife Emina: I love you! Thank you Mats Remman for teaching me about computers for the last thirty years, and more recently how to make sense of actual computer code.

A large and final thank you to my friend Thomas Rognmo. The feedback and encouragement you gave me in the recording studio when we were producing the audio prompts for the *MinTest* application was invaluable. Beyond this, your creativity and intelligence inspired me through all the years we knew each other. I miss you and I deeply regret that I did not do more to help you in the time you were among us.

Terje Bektesevic Holmlund, Tromsø, 04.10.19

2 Abbreviations

ASR	Automatic Speech Recognition
ANOVA	ANalysis Of VAriance
BERT	Bidirectional Encoder Representations from Transformers
LSA	Latent semantic analysis
MATRICS	Measurement and Treatment Research to Improve Cognition in Schizophrenia
ms	Milliseconds
NIMH	National Institute of Mental Health
NLP	Natural Language Processing
RDoC	Research Domain Criteria
rmANOVA	Repeated-measures analysis of variance
RT	Response time
s	Seconds
SD	Standard Deviation
USA	United States of America

3 Abstract

Detecting signs of disorder from listening to spoken words is a core method in psychiatry. Traditionally the interpretation of speech depends on inherently subjective processes. By contrast, digital technology can be leveraged to detect and analyze what words are spoken, timestamp when they are uttered and quantify the manner in which they are expressed. With the use of mobile communication technology, digital speech processing tools are possible to use outside of traditional laboratory settings. This thesis argues that the necessary infrastructure to move speech processing into clinical practice is currently available. To examine this claim, a mobile application for remote mental state assessments was developed that implemented speech-based neuropsychological testing in 353 participants in two countries. It was possible to collect speech data in ecologically valid settings, but future larger scale implementations must solve technical, legal and cultural challenges by interdisciplinary teamwork. The findings of spoken responses on the classic Stroop color-word test from 57 patients with substance use disorders and 86 healthy participants showed that the production of single-word speech utterances could be measured with a high level of temporal precision. The classic Stroop task response latency interference was replicated and the scope of measurements was extended with novel speech characteristics. The audio files from 59 participants naming words in a category fluency task could be analyzed for both the temporal dynamics of response-word sequences and the semantic relatedness between words. Finally, the story recall ability in 25 patients with serious mental illness and 79 healthy participants was examined, and automated measurements of their ability to retell a story was computed using both simple word-count procedures and more advanced estimates of distances in a semantic vector space. In conclusion, it is technologically feasible to develop instruments for measuring multiple aspects of how patients with psychiatric disorders speak, and traditional speech-based neuropsychological tests can be employed outside of a laboratory setting provided the digital infrastructure is able to ensure the privacy of the users.

4 Sammenfatning på norsk (Norwegian summary)

Tittel: Modellering av talemateriale som er samlet inn ved hjelp av stedsuavhengige tjenester: Nytteverdi for psykiatri

Det å kunne oppdage tegn på sykdom hos noen ved å lytte på hvordan de snakker er et viktig klinisk verktøy i psykiatri, men tradisjonelt sett har dette vært avhengig av svært subjektive vurderinger hos den som lytter. Digital teknologi kan være behjelpelig i slike vurderinger ved å fange opp hvilke ord som blir sagt, tidfeste dem, og tallfeste forhold med måten de blir uttalt på. I tillegg åpner digitale løsninger opp for at man kan gjøre slike analyser ved hjelp av mobiltelefoner eller annet mobilt utstyr, såkalte stedsuavhengige tjenester. Denne avhandlingen hevder at den digitale infrastrukturen som trengs for å benytte slike verktøy til kliniske formål nå er tilgjengelig. For å undersøke om dette stemmer ble det utviklet en mobilapplikasjon for gjennomføring av tale-basert nevropsykologisk testing og denne ble brukt av til sammen 353 forskningsdeltakere i to ulike land. Det var mulig å samle inn talemateriale uavhengig av hvor deltakerne befant seg, men både tekniske, juridiske og kulturelle utfordringer måtte løses gjennom tverrfaglige samarbeid for å få effektiv bruk av de nye metodene. Den første artikkelen i avhandlingen dreier seg om disse utfordringene.

Avhandlingen forteller også om tre ulike tester som viste seg å gi nyttige målinger av hvordan deltakerne snakket. Hver av disse testene har blitt viet en egen artikkel. De tre testene illustrerer hvordan det er mulig å gjøre beregningsbaserte analyser tale på tre ulike nivåer, nemlig for enkeltord som blir sagt, for flere ord når de settes sammen, og for komplekse ytringer som skal ses i sammenheng med en kontekst. Ved å undersøke tale fra den klassiske Stroop-testen var det mulig å tidfeste svært nøyaktig når deltakerne responderte på oppgavene. I denne testen kom ulike ord opp på skjermen til deltakerne (57 pasienter med rus- og avhengighetslidelser og 86 friske frivillige), og oppgaven var å si hvilken farge ordene var skrevet i så fort som mulig. Analyse av enkeltordene som ble sagt kunne avdekke den klassiske Stroop-effekten, altså at betydningen til ordene som kom opp på skjermen påvirket hurtigheten til navngivingen av fargene. Graden av denne påvirkningen kan brukes til å måle evne til å holde oppmerksomhet rettet til en oppgave, noe som kan være en del av utredning og oppfølging av mental helse. En annen oppgave testet evnen til å holde verbal flyt. Deltakerne (24 pasienter med rus- og avhengighetslidelse og 35 friske frivillige) fikk

oppgaver hvor de hadde ett minutt til rådighet for å si så mange dyre-ord som de kunne komme på, så fort som mulig. Lydopptak av svarene som ble samlet inn kunne gi ny informasjon om tidsmessige forhold i slike serier med ord, og i tillegg kunne objektive metoder brukes for å beskrive sammenhenger i meningsinnholdet i ordene som ble brukt. Evne til å holde god verbal flyt kan være en del av utredninger som avdekker både psykiatriske og nevrologiske sykdommer. Til slutt beskrives en metode for å gjøre en automatisert måling av hvor godt deltakerne kunne huske historier de ble fortalt. Ulike historier ble avspilt fra mobilapplikasjonen og deltakerne (25 pasienter med alvorlig psykisk sykdom og 79 friske frivillige) hadde som oppgave å gjenfortelle dem så godt som mulig. Automatisk talegjenkjenning kunne gjøre lydopptak om til tekst, og ved å bruke språk-teknologiske metoder var det mulig å tallfeste likheten mellom den originale historien og gjenfortellingen på en slik måte at det samsvarte godt med den måten menneskelig personell vurderte gjenfortellingene. Tap av hukommelsesfunksjon kan være et viktig tegn på sykdom, derfor kan slik testing benyttes som en effektiv del av oppfølgingen av hjernefunksjon og mental helse.

Funnene i forskningsprosjektet viser at det var mulig å gjøre nøyaktige analyser av måten deltakerne snakket på selv om registreringene var gjort via en mobilapplikasjon. Det at oppgavene som ble brukt kunne si noe om viktige funksjoner som oppmerksomhet og hukommelse styrker muligheten for at denne nye måten å undersøke snakking på kan være nyttig. Målemetodene som er presentert kan utvikles videre til å gi en bred og utfyllende beskrivelse av måten pasienter uttrykker seg på. Gjentatte og systematiske registreringer av stemmebruk kan i tillegg gi et verktøy som beskriver hvordan mentale tilstander endrer seg over tid. En forutsetning for at disse nye metodene kan lykkes er at datasystemene opprettholder et sterkt personvern hvor den enkelte har god kontroll over sine egne data. Hvis denne informasjonen kan gjøres lett tilgjengelig for både klinikere, pasienter og forskere så kan den potensielt både bidra til bedre behandling og bedre vitenskapelig forståelse av mentale tilstander.

5 List of papers

I.

Holmlund, T. B., Foltz, P. W., Cohen, A.S., Johansen, H., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological Assessment, 31*(3), 292-303. doi: 10.1037/pas0000647

II.

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S., Bernstein, J., Rosenfeld, E., Laeng, B., & Elvevåg, B. (submitted). Using automated speech processing for repeated measurements of attentional bias and control. Manuscript submitted for publication.

III.

Holmlund, T. B., Cheng, J.,Foltz, P. W., Cohen, A. S., & Elvevåg, B. (2019). Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry Research. 273*, 767-769. doi: 10.1016/j.psychres.2019.02.014

IV.

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., D., Cheng, J., Bernstein, J., Rosenfeld, E., & Elvevåg, B. (submitted). Applying speech technologies to assess verbal memory in patients with serious mental illness. Manuscript submitted for publication.

6 Introduction

Spoken words convey crucial information regarding the health of humans. This is an important premise for psychiatry, because by listening to the words spoken by patients clinicians can discover important clues about the mental states of the speakers. Although the subjective *symptoms* sometimes described by patients in clinical conversations may provide important clues about pathological processes, this thesis concerns itself with measuring *signs*, namely the observations clinicians can make regarding external, measurable states of the speaker. The distinction between signs and symptoms is important. Symptoms are *experienced*, like the feeling of vertigo when standing on a high and steep cliff. Signs are *observed*, such as when an electrocardiogram records an elevated rate of heartbeats before a dreaded surgical procedure. Interpreting the signs in verbal behavior can provide a unique window into thought processes (Elvevåg et al., 2017), and reliable measurements of disordered speech and language may be a useful tool for understanding psychiatric disorders. This thesis will claim that technology has matured to the point that it is possible to effectively and remotely collect speech data outside of traditionally controlled laboratory settings and that this affords fast computational analysis of verbal behavior, and that such analyses can provide information critical to the assessment of attention, verbal fluency and memory functions. If this claim is true, it should be possible to collect data on spoken words and learn valuable lessons that can serve as the foundation for constructing better models of human behavior and ultimately for the future development of clinical tools in psychiatry.

Models represent expectations of our surroundings based on experiences of the world, and in the case of descriptive scientific models, they are based on measurement data. Consider a simple everyday example of a model, namely a clay figure in the likeness of a giraffe. We find it has four sticks to on the ground holding up a barrel-like blob of clay and we interpret this to be an approximation of four legs and a torso. If this is a good model of reality, based upon our experience we will then expect there to be a long neck connecting the head to the rest of the model. To put it in the context of a model of verbal behavior, one can consider experiences of hearing spoken words. If someone speaks with a calm voice, one would expect them to continue to do so for the duration of the conversation, not suddenly to switch to a voice booming at two hundred decibels (which is about the sound of NASAs Saturn V rocket!). Put differently, scientific models of human speech can represent such expectations,

and such expectations are the basis of much technology that society depend on today, such as automatic speech recognition. As with the example of the clay giraffe, the level of detail in models matters. A lump of clay with stick legs can probably be recognized as a giraffe as long as the characteristically long neck is present, but a model with carefully crafted detail of bones, muscles and fur will probably be more useful in the training of veterinary scientists. This is also the case with models of speech and language, as the scientific or clinical utility can increase with higher resolution (Cohen et al, in press). Indeed, a balance of both theory and data must inform models, and if there is a lack of data-driven reasoning then crucial aspects of the modelled phenomenon may be missed (Silvert, 2001). Currently, some relevant aspects of speech data are hard to acquire due to technical and legal challenges. Therefore, the results presented in this thesis are not intended to provide new data for the descriptive models of verbal behavior, but rather they aim to demonstrate what is possible in terms of leveraging new technology for data collection.

Robust and precise measurements of overt behavior can provide an important addition to an increasingly detailed understanding of human behavior in general, laying the foundation for an integrated scientific understanding of mental states across several levels of analysis. Research in psychiatry can also be situated within several levels of description. At one end of the spectrum, it is possible to investigate phenomenological descriptions of symptoms or "*lived experiences"* via self-reports from patients (e.g., personality disorders; Shepherd, Sanders, & Shaw, 2017), while at the other end of the spectrum it is possible to link minute details of molecular mechanisms in psychiatric disorders (e.g., depression; Fox & Lobo, 2019). In recent years, research in psychiatry has had a strong focus on levels of descriptions that lie close to the neurobiological level, but these may not be sufficient to create meaningful explanations of complex behavior (Krakauer, Ghazanfar, Gomez-Marin, Maciver, & Poeppel, 2017). In this thesis approaches are discussed that may serve as useful additions to the toolkit of both researchers and clinicians at this intermediate level, namely the quantitative descriptions of complex verbal behavior. The successful collection and analysis of numerical speech data is therefore important because it can be subjected to mathematical modelling and ultimately prediction of states of mental health.

Conveniently, the sounds of speech are easily available for measurement using microphones, and computational analysis is possible given models that describe the statistical patterns that

can be derived by examining previously recorded language expressions. Speech is neatly organized into behavioral units of time-series, the words, and these units can be arranged in temporal structures as sentences. This organization has been under development in human societies spanning millennia, underlying effective means of conveying information, creating a system of expected connections between symbols under the overarching term "language". These expectations constitute language models. For an illustration of how models of language are internalized, consider hearing someone say the following and abruptly stopping the utterance before finishing: "*I am thirsty, so I am going to get a glass of ... "*. In this case, a model of language expressions allows us to predict what should come next, perhaps giving us an expectation of the words "*water*" or "*milk*" (Kuperberg & Jaeger, 2016). Importantly in psychiatric settings, a sense of surprise can be generated in cases where speech violates expectations. If the speaker uttered "*... I am going to get a glass of flames*", the meaning of the utterance would be much less clear to us, perhaps pointing towards a sign of a disordered mental state (for a two-part review, see Kuperberg, 2010a; 2010b). These language expectations can be quantified in computational models built on large amounts of language data. The example of leaving out (i.e., "masking") a word from a sentence and making guesses (based upon expectations) is both relevant to the very large number of studies that have used the half-century old Cloze procedure (Taylor, 1953), but the approach is also part of the state-of-the-art methods for training computational language models such as Bidirectional Encoder Representations from Transformers (BERT: Devlin, Chang, Lee, & Toutanova, 2019). Language models provide the basis for more structured and robust descriptions of verbal behavior, ultimately enabling an entirely new representation of signs of disordered mental states.

While the importance of compassionate human helpers in the clinic cannot be overstated, reliable records are needed to make repeated measurements of mental states, and as such irrelevant states of the observer (i.e., clinician) should have the least amount of effect on assessment. Ultimately, reliable methods of measurements will also lead to more reproducible investigations, better models of evolving patterns of behavior and therefore be of great value to scientific and clinical endeavours. This thesis states that technology has reached a point of maturity to enable useful data collection outside a scientific laboratory. The focus now narrows to how speech and language has been considered in psychiatry, exemplified by how it has been conceptualized in the case of schizophrenia specifically. After this, there will be a

specific emphasis on speech-based assessment of specific neuropsychological functions crucial to mental states, notably attention, verbal fluency and memory. Ultimately the nature of how new technological solutions has started to transform psychiatry is reviewed, before descriptions of the aims, methods and results of research program that constitutes this thesis.

6.1 Speech and language in psychiatry

Listening to spoken words provides the basis for mental status examinations in psychiatry, where patients express words describing their symptoms and clinicians listen while observing for behavioral signs of disease. The departure point in the discussion of speech and language in psychiatry is the neurodevelopmental disorder schizophrenia, where disordered speech is one of the key criteria for diagnosing the disorder (American Psychiatric Association, 2013). Indeed, the presence of language disturbances as a presenting sign of schizophrenia has been noted in the scientific literature for over a century (e.g., Bleuler, 1911). There are some overarching characteristics of the disorder that have been found to have comparable incidence in human populations around the globe and so it has been suggested that some of the core signs of schizophrenia are likely intrinsic to the *Homo sapiens* (Crow, 1998). The consistency across populations, even if they likely had been separated for ten millennia (Jablensky et al., 1992), points towards a putative biological basis to the signs of some disordered mental states, something that in turn may be modelled and understood given proper measure assays of the aberrant verbal behavior.

Clinicians can intuitively form an opinion regarding when something is common or unusual behavior in a specific context based upon their model of behavior, with the observations being either surprising or expected. Speech that is not in agreement with such expectations may result in ineffective communication and lead to a suspicion that a disordered mental state is present. Experienced clinicians have collected and described patterns of speech, and developed formal tools of assessment. One of these tools is specifically relevant to this discussion, namely the scale for the assessment of Thought, Language and Communication (TLC) developed by Andreasen (1986; Andreasen & Grove, 1986). On this scale, verbal communication can be rated on eighteen different categories such as "*poverty of speech*", "*tangentiality*", "*incoherence*" or "*perseveration*". Some of these categories seem measurable

only by subjective judgement of an observer, while other categories afford more objective measurements, such as in the case of “pressured speech” where a rate of over 150 words per minute is noted as rapid or pressured. This serves as an illustration of how signs that seem dependent on subjective judgement nonetheless may be amenable to numerical operationalizations by means of tools that can measure speech production. Although the concept of thought disorder that this scale was designed to assess can be problematic since the term is so wide (Andreasen, 1982), the decades old framework is still highly influential receiving frequent citations in contemporary literature. Building on the knowledge gained through tools such as the Thought, Language and Communication scale it is possible to construct operationalizations of disordered speech and adapt them to a computational environment.

In psychiatry, the definition of unusual verbal behavior is complex as it can include highly heterogeneous behavior and so for assessment purposes there is a benefit in narrowing down the specific tasks that elicit the speech to be analyzed. For example, compared to more discourse-like interviews, techniques for neuropsychological assessment can employ a constrained questions-and-answer approach, providing tasks to be solved by the patient. More constrained tasks has the added benefit that it is possible to assess more specifically the effect psychiatric disorders can have on neurocognitive function. After an extensive review of the literature, DeLisi (2001) concluded that while there was no clear pattern in speech characteristics across studies of patients with schizophrenia, the majority of observations could be associated with disturbances in attentional processes or working memory. Indeed, in the case of schizophrenia, it is well established that impairment of cognitive functions lie at the very core of the disorder itself (Elvevåg & Goldberg, 2000; Kuperberg & Heckers, 2000). The centrality of cognitive function to the illness is why a new approach to assessment instruments was developed to assay neurocognition for clinical trials in schizophrenia. This approach was based upon a consensus from very many clinical research groups on what actually constitutes the most valid and reliable assessment method of the core underlying neurocognition (e.g., the US National Institute of Mental Health (NIMH) Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS); Marder, 2006; Kern, Green, Nuechterlein, & Deng, 2005). This marked a dramatic shift in the way psychiatric disorders were conceptualized and was a crucial building block in the subsequent development of the National Institute of Mental Health Research Domain Criteria (NIMH

RDoC; Insel et al., 2010), which unlike the gold-standard diagnostic manuals in psychiatry (i.e., Diagnostic and Statistical Manual of Mental Disorders and International Classification of Diseases) enabled the conceptualization of psychiatric illness in terms of different types of behavior (e.g., attention, language, memory). The NIMH RDoC also provides a research framework to deliver the promises of precision medicine in psychiatry by targeting cognition (Insel et al., 2010). Therefore, this thesis sought to assay cognition with tasks tailored to probe specific cognitive processes, notably attention, memory, and language, in order to develop the next generation tools to assay mental states with increased precision and clinical translation value.

6.2 Neuropsychological assessment using speech responses

For the purpose of this thesis, the focus is on three tasks with a long history in neuropsychological assessment and which have spoken responses, namely the Stroop task for assessing attentional control, a category naming task for assessing verbal fluency and a story retelling task for verbal memory assessment. Task such as these can serve as proxies for more complicated real-world behavior. By investigating the performance of different individuals on these “models” of situations that can be encountered in the real world it is possible to actually build better models of human behavior. The administration procedures of neuropsychological tasks have already been transformed after the introduction of digital technology, increasing the opportunities of what can be effectively investigated in the laboratory with easy-to-use software for stimulus presentation and response collection (e.g., E-Prime; Psychology Software Tools, Pittsburg, PA). This thesis will argue that these tasks are also possible to adapt to a format where they can be used outside of the traditional psychological assessment laboratory, where it is possible to make measurements of higher ecological validity. Going beyond merely digitalizing the administration of traditional neuropsychological test, there are also possibilities to derive magnitudes more information from recorded speech responses than can be captured with the traditional use of stopwatches, pencils and paper.

6.2.1 The Stroop Color Word test and measurements of attentional bias and control.

One of the oldest established tasks of neuropsychological assessment is the Stroop task (Stroop, 1935; MacLeod, 1991). In this task, people are commonly asked to name the color that different words are printed in, as fast as possible, and avoid being distracted by the content of the words (e.g., if the word RED is printed in blue ink, the meaning of the word "red" should be ignored and the correct response is "blue". When responses are spoken, this task leverages measurements of *hesitations* before single words are spoken in order to assay biases in attention and the ability to control and override such biases. Everyday life presents a complicated world to navigate, and the ability to selectively engage events in the environment with adaptive behavior and cognition are core to the successful proliferation of the human species. This ability to manage cognitive resources for task-related behavior can allow humans to override habitual, impulsive and short-sighted tendencies (Shenhav et al., 2017), and adapt to specific tasks by adjusting perceptual selection, response biases, and on-line maintenance of contextual information (Botvinick et al., 2001). Difficulties in engaging in adaptive behavior, be that complex or simple, are likely at the core of a variety of behavioral patterns observed in disorders that affect cerebral function (e.g., in schizophrenia (Green, 1996) and in major depressive disorder (Diener et al., 2012). These difficulties have been the primary target of measurement for traditional clinical Stroop tasks (e.g., the Golden Stroop; Golden, 1976).

Methods for administering Stroop tests have been under continuous development, and interestingly the literature comparing older methods, (i.e., presenting lists of words on printed cards, the Card Stroop) and newer paradigms (i.e., presenting single words on computer screens, the Single Trial Stroop) have found that differences between healthy individuals and patients tend to be smaller when employing the newer methods (specifically for schizophrenia: Westerhausen, Kompus, & Hugdahl, 2011). Possible reasons for this may lie in several aspects of the procedures and choice of stimulus material, such that an effective and adaptable single-trial approach may be superior in terms of developing objective and robust measures to characterise the severity and profile of attentional biases and control deficits. There have been attempts at using the Stroop task for remote assessment via mobile devices (Pal et al., 2016; Allampati et al., 2016), but these have had the disadvantage of requiring

touch-screen responses for responses, thereby introducing a more complex motor response and technical challenges for precise temporal measurements (i.e., dependent on the refresh- and sampling rate of the screen). Using spoken responses on this task provides a well-validated and convenient way of collecting a rich set of data on individual performance.

The single spoken word is at the first level of computational analysis of speech data: Is it possible to measure *when*, *what* and *how* a word is uttered with sufficient precision for it to be useful? Traditionally, the slight "hesitation" in color naming if there is a conflict between the letter color and the written word (e.g., "PURPLE" written in a red color) has been measured using laboratory microphones to obtain millisecond precise timing information on the delay between the appearance of a stimulus appears and the voice level reaching a certain threshold. With more detailed analysis of speech recordings it should be possible to measure the durations of the word utterances, the corrections (e.g., the utterance with hesitation such as "*Gree...no red!"*) and acoustic features such as pitch. This type of information would provide the fundamental building blocks at the single word level for understanding the dynamic process of speech responses and corrections thereof. In the near future, it is possible to envisage that a complex combination of acoustic features and duration of output would be useful to differentiate clinical conditions (e.g., a slow, flat, long utterance from someone who is either uninterested or depressed, versus a fast, high pitch and quick energetic verbal response in someone who is very engaged or manic). Differentiating between complex psychiatric conditions and sheer lack of interest in the task or lack of willingness to participate will be challenging, but possible given proper data collection and analytical frameworks.

6.2.2 Verbal fluency tests can provide measurements of the flow and timing of multiple spoken words.

Another widely used language task in psychiatric research is the category verbal fluency task. A key factor in its success is probably that it is easy to administer and takes little time to complete. In the task, participants are given a noun category (e.g., animals, vegetables) and asked to verbally produce as many examples of nouns in a specific category for a specified duration (e.g., one minute). The experimenter writes down the response words and assigns a point for each unique examplar produced. Fewer responses have been associated with a wide

variety of cerebral pathologies ranging from structural brain lesions to diffuse brain injuries and dementia (Lezak, 2004). Verbal fluency has been shown to be dependent on frontal and parietal regions of the left cerebral hemisphere (Friston, Firth, Liddle, & Frackowiak, 1991), providing a useful and easily deployed tool for assessing cortical health. While this test has provided useful measurements in research areas such serious mental illness (e.g., schizophrenia; Bokat & Goldberg, 2003; Nicodemus et al., 2014) and dementia (e.g., Henry, Crawford, & Phillips, 2004), the traditional operationalization of simply counting the number of words produced ignores the obvious fact that even on such a simple task there is a remarkable amount of structure and temporal information (Bousfield & Sedgewick, 1944; Bousfield, Sedgewick, & Cohen, 1954). Therefore, simple counts alone run the risk of potentially missing fundamental aspects of the dynamical component of language production (Elvevåg et al., 2016). Indeed, when naming words from a specific category (e.g., animals) over the course of a minute, the dynamical aspects of the retrieval process are missed if performance is scored as only the total sum of words uttered. What knowledge might be gained if it were possible to assay these dynamical aspects? Indeed, if detailed temporal information is available, this can provide a window into how speech production is initiated and maintained, as suggested from a finding of slower speech within defined time periods observed in patients with serious mental illness (Krukow, Harciarek, Morylowska-Topolska, Karakuła-Juchnowicz, & Jonak, 2017).

In addition to these temporal aspects, semantic relationships between the words within the noun categories used can also affect speech production. For example, within the category of “animals”, dependent upon cultural experience, it might be expected that there is a stronger semantic relationship between different types of farm animals (e.g., cows, sheep, chickens) than between farm animals and creatures from the African savanna (e.g., zebras, giraffes and gazelles). Therefore, by measuring the patterns of semantic similarities between successive words it might be possible to assess the efficiency of the memory search processes within putative networks of semantic entities. Such measurements depend on effective quantification of semantic similarity between words, and recent advances in this area have been made employing computational language models (Pakhomov, Eberly, & Knopman, 2016; Kim, Kim, Wolters, MacPherson, & Park, 2019).

A key notion in understanding why language can be perceived as incoherent is this: If the succession of words are not confined to a topic or meaningfully linked, they can be hard to understand. Therefore, large and unusual transitions in semantic relationships can be perceived as less coherent. This phenomenon might be intertwined with temporal dynamics of speech but the interaction of dynamics (time) and coherence has never been previously examined in this task. For example, an effective recollection that seemingly systematically retrieves "farm" words such as *cows -sheep -chickens* might be faster (i.e., shorter inter word intervals between these particular words) than when retrieval switches between putative sub-categories such *chickens* in the aforementioned sequence being followed by an animal from the African savanna such as *zebras.*

6.2.3 Retelling a story can enable measurements of verbal memory ability

Another key aspect in how the content of speech might reveal signs of disordered thinking is how connected it is to the context or discourse where it appears. One way to assess the ability to produce speech that is connected to a given context is to measure the ability to retell a story. An important clue as to why this is relevant to the assessment of mental states comes from studies involving patients diagnosed with schizophrenia. In these patients, it has been shown that the verbal memory processing component of remembering events (i.e., episodic memory) can be specifically impaired, while more visual aspects of processing are not (Aleman, Hijman, de Haan, & Kahn, 1999; Cirillo, & Seidman, 2003). Therefore it has been suggested that problems with verbal memory may serve as a useful endophenotype for the disorder (Skelley, Goldberg, Egan, Weinberger, & Gold, 2008).

Verbal memory assessment has traditionally been very resource-demanding. Personnel are trained in administering testing procedures, with testing material that is typically under strict copyright protection (e.g., Wechsler Memory Scale; Wechsler, 1997). There are good reasons for this, namely to ensure that assessment is well validated and can be trusted. However, it does make frequent testing infeasible. While verbal memory ability is assumed to be a rather stable cognitive ability, there is also the possibility that measurements repeated on frequent and regular schedules may reveal new information about clinical states and effects of treatments. It stands to reason that being able to assess something so fundamental on a

frequent basis will be more accurate than cross-sectional “snap shots”, notably because verbal memory is likely to be affected by clinical state changes. Thus, to be able to accurately, objectively and reliably assay this promises to provide the much needed therapeutic target for monitoring treatment responsiveness,

The goal would be to be able to objectively, reliably, frequently, and automatically measure if someone is speaking about an expected topic, namely re-telling a story. To achieve this, this research program sought to create an automated method of rating participant recalls that performed as well as humans. If successful this would afford an assay that is clinically sensitive (as it can be obtained frequently), reliable (as it is objective and rated by machines) and critically it could serve as the much needed clinical end point to gauge treatment responsiveness and clinical state.

6.3 Two core analytical methods for analysis of speech production and semantic content

Computer software has been assisting in the analysis of spoken words progressively for decades, and it is now possible to effectively transform the sound of speech into lexical representations (i.e., transcribe sound to words), and transform lexical representations into numerical representations (i.e., word vectors). In recent years there has also been a democratization of these software tools through the widespread sharing of computer code (i.e., open source-code software). For the first step, automatic speech recognition software such as Kaldi (Povey et al., 2011) is available free of charge for researchers who can deal with the intricacies of managing the software and build acoustic and language models from their own data. For the second step, methods to analyse transcribed text using language models based on what is known as word vector embeddings is available, equally free of charge and with effective software packages for implementation. These two core approaches, namely Automatic Speech Recognition (ASR) and word vector methods will now be described in greater detail, as they were employed in the current research program which sought to develop precise and robust measurements for remote assessment of mental functions for use in psychiatry.

6.3.1 Automatic Speech Recognition for assessing speech production.

The digital processing of recordings provides an opportunity to measure speech production in detail, and analyzing *when*, *what*, and *how* something is said can reveal important clues about the person from whom the sound originated. Underlying speech are a complex series of behavioral elements, including the contraction of muscles of the thorax to force air out of the respiratory tract and fine-tuned activation of the laryngeal muscles to adjust the firmness and position of the vocal cords. This generates the sinusoidal fluctuations of air pressure, namely sound, which can be subjected to speech signal analysis when captured by microphones.

At a high level of abstraction, speech recognition can be conceptualized as a classification task, where the objective is to determine if a certain part of the sound-wave belongs to an uttered word or not. The one-dimensional signal of sound is decomposed into time-frequency information using Fourier transformations, most commonly transformed into Mel-frequency cepstral components. The “mel” part has roots in psychophysics, where the scale is more akin to how the human auditory system perceives differences in frequencies. After these features have been extracted, a typical approach is to slide a 25ms window across the timeline in 10ms steps and compute the probabilities of the presence of certain phonemes within the windows, in part using hidden Markov models. The step size of 10 ms is important, as it defines the 10ms temporal resolution of the timestamping methods described in this thesis. Detected phonemes are then processed according to a lexicon to compute the likelihood that a certain word has been spoken. How this lexicon is defined and created can allow for customization towards specific categories of words, for example an animal word or a color word, increasing the accuracy and utility in certain specific cases. However, such customization will come at a cost, as the accuracy in detecting more general language expressions will be lower. Ultimately, the most likely lexical items are then processed with regards to a language model that adjusts the output words based on likely sequences of words (Young et al., 2006).

Speech recognition technology is very mature and the techniques may inform on ways to analyze other time-series data of behavior. The methods for detecting words in sound recordings represent an example of a core approach in machine learning, namely that a slice of data (in this case a 25ms section of a sound recording) is classified as belonging to a specific category, namely either representing a “silent” part of the recording (<SIL> in Figure

2, panel E) or representing one of a specific number of different phonemes that ultimately constitutes words. The manner in which the phonemes are then placed within a larger context (i.e., words), and that the sequence of words can be amenable to modelling, reveals that what is expected can be expressed across several levels of verbal behavior. In a similar way to how one can understand that another motor action, such as an extending arm, can be a part of higher level action such as a handshake, these methods provide ways to analyze behavior as it unfolds, employed to a signal that is in its very nature inherently sequential and dynamic (Deng & Li, 2013). This is encouraging because the mathematical foundation for the signal processing and word detection mechanisms are very mature, and it stands to reason that this type of framework can inform on other domains where time series data are to be analysed and events detected and classified. This is important to emphasize as it is possible to create mathematical models of behavior from the very small scale, namely vibrating vocal cords to create identifiable phonemes, to the more complex, such as retelling a story previously heard.

6.3.2 Natural Language Processing and word vectors for semantic analysis

Natural language processing, commonly abbreviated as “NLP”, refers to computational methods that are leveraged to understand, analyse or even alter digital records of language expressions as they occur in natural settings, most commonly in the form of written words. The methods can be employed for a variety of purposes, for example to search for and extract specific information in large volumes of text, to translate text automatically from one language to another, or to get a score of the sentiment that is being expressed (e.g., if a restaurant review is positive or negative). The boundaries of the concept of NLP are not clearly defined, and methods like automatic speech recognition is sometimes included in this umbrella term. In terms of academic disciplines, NLP draws on developments in computer science, statistics and linguistics. For the purpose of this thesis, methods are employed to derive quantitative measures of the semantic content of utterances, to analyze the *meaning* of what has been said in a setting of mental state assessment.

The meaning of a word can be represented as a numerical vector derived from analyzing how it co-occurs with other words in common use in language. Modern language processing methods adopt a very practical approach to defining the meaning of a word, and this approach

is in line with a famous statement from the highly influential language philosopher Ludwig Wittgenstein, namely: "*The meaning of a word is its use in the language*" (Wittgenstein & Anscombe, 1967, §43). A little later the linguist John Rupert Firth (1957, page 11) stated: "*You shall know a word by the company it keeps!*". This line of thinking has become formalized in what is known as the "Distributional hypothesis" in linguistics, namely that linguistic items that appear in similar contexts, or show similar distributions in their use in the language, tend to have similar meanings (McDonald & Ramscar, 2001). What this means in practical terms is that statistical language models can be built from analysing how words are used in natural language expressions, most commonly collected as large corpuses of text, for example books or samples of internet pages. Given a sufficiently large corpus of words, it is possible to apply mathematical computations to define words based on the context where they commonly occur, stated differently: *"the company they keep"*. Some of the practical ways with which researchers have approached this previously will be described. This description attempts to avoid the notion that computations around the "meaning" of words are some kind of black box that magically discovers some unfathomable truth about nature: Computational semantic analysis is ordinary mathematics based on common usage of words.

An early mathematical manifestation of these ideas is Latent Semantic Analysis, which was patented in 1988 by Thomas Landauer and colleagues (1997). The researchers had collected 30,000 text samples with a length of approximately 150 words. In total, there were 60,000 different unique words that were organized into a matrix: One row for each word, one column for each text sample, namely "*the company it keeps*". Each cell in the matrix represented a count of how many times a word occurred in a particular text sample. The matrix was then condensed from 30,000 to 300 dimensions using a method called Singular Value Decomposition, a method somewhat akin to factor analysis. This method thus provided each word with a vector that described its size and direction in a semantic space, and the word vectors could be said to be "embedded" in this space.

Another way to describe word embedding spaces is that they are geometrical spaces that a word will be "located" in, somewhat comparable to the location of a city on a two-dimensional map, as defined by two numbers namely longitude and latitude. The high-dimensional space is impossible to visualize properly, as 300 dimensions go well beyond the three or four dimensions that are easy to conceptualize. However, two-dimensional

simplifications, like the one in Figure 1, panel E, enable visualization of the concept. For example, words such as “green” and “red” can be embedded quite close in a semantic space, as they often appear together in descriptions of colors of objects, while the words “knife” and “fork” may be embedded together but some distance away from the cluster of color words. The angle between the vectors for “knife” and “fork” is small, and this is often expressed as the cosine similarity, sometimes as the Euclidean distance, with small distances equating to related or similar meanings, in this case that they belong to the same category of utensils.

Since the introduction of semantic embedding models over thirty years ago, progress has been made both in terms of the algorithms that compute the vectors, as well as the possibility of leveraging increasingly large text corpuses to improve performance (Bengio et al., 2003). Building on the conceptual foundation from early implementations of Latent Semantic Analysis, neural network approaches such as the Word2vec algorithm (Mikolov, Chen, Corrado, & Dean, 2013) and the global log-bilinear regression algorithm Global Vectors for Word Representation (GloVe; Pennington, Socher, & Manning, 2014) have displayed improved performance. Several companies have further developed these types of methods, among the notable recent advances are the deep Bidirectional Transformers for Language Understanding (BERT: Devlin, Chang, Lee, & Toutanova, 2019) and XLNet (Yang et al., 2019) from Google, as well as the second Generative Pre-training model from the non-profit organisation OpenAI (GPT-2; Radford, Wu, Child, Luan, Amodei, & Sutskever, 2019). In this multitude of different approaches there is yet to emerge a single definitive “best practice” and there will likely be several methods that will be sufficiently useful, depending upon the purpose of the processing.

In the current evaluation of new technologies for computational analysis of remotely collected speech data, two different but related measures of semantic similarity are employed, namely the aforementioned cosine similarity which is excellent for computing word-to-word similarity in a verbal fluency task (using GloVe-embeddings), and Word Mover’s Distance (Kusner, Sun, Kolkin, & Weinberger, 2015) which is more suited to assessing the overall distance in semantic content between two multi-word entities of recorded language. The two methods described, namely speech recognition and semantic vector space analysis, can be useful to detect Stroop hesitations and measuring decreased verbal fluency or problems in memory recall. These types of neuropsychological assessment are commonly conducted in

controlled clinical or laboratory settings, with humans listening to responses, making notes with a pen and paper. Technology can be employed to provide additional information that may be valuable in assessment, if it is accepted by both clinicians and patients. Acceptability is a key issue in how the spoken words that are uttered in these types of testing procedures can be utilized. However, to get such quantitative information that may be of value in predicting mental states new digital tools are needed for collection, transfer, processing and storage of information into the healthcare system.

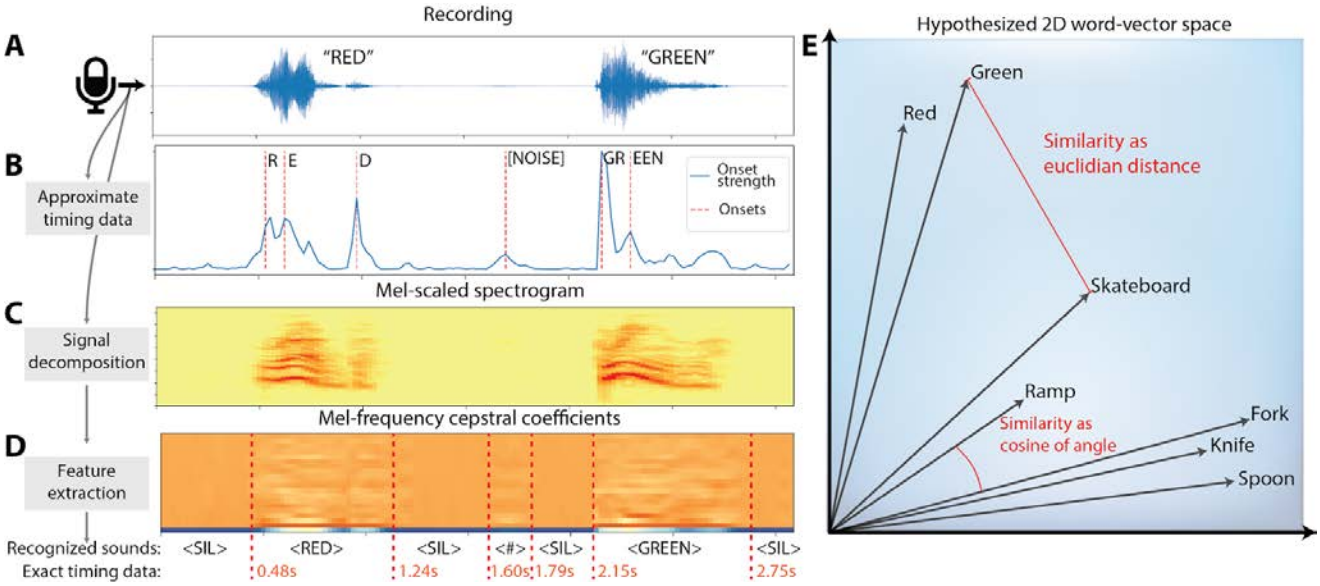


Figure 1.

Two essential tools for computational analysis of spoken words in psychiatry, namely speech recognition processing and semantic vector space models. Panel A: Analysis starts with processing digital recordings of speech. In this example the speaker says the two words "RED" and "GREEN". A familiar way to plot the information is as a function of the amplitude of the sound pressure waves over time. Panel B: By analysing the dynamics of the peaks of the sound pressure waves, it is possible to get rough estimates of the onsets of spoken phonemes as well as non-speech noises. Panel C: More detailed information can be gained by decomposing the speech signals into its constituent frequencies, here represented by an example of a time-frequency plot, or spectrogram. The y-axis represents frequencies of the sound, such that higher pitched sounds are represented by darker colors higher in the plot. In this case, the y-axis is mel-scaled, meaning that differences along the y-axis are more representative to how humans perceive differences in sound frequencies. Panel D: The time-frequency information is transformed into numerical features that are optimized to create acoustic models of what sound-waves correspond to phonemes and words, in this case what is called mel-frequency cepstral coefficients. Based on these features, the automatic speech recognition models can output the words that are likely to have been spoken, and also the temporal information about when the words occurred, with a +-10 ms temporal resolution. Panel E: By processing large corpuses of text, often on the order of several hundred million words, it is possible to construct spatial representations of semantic relationships. Words that often co-occur in natural use of language will be embedded close in such semantic space, exemplified here by utensils such as "knife" and "fork" being located close together, but not with unrelated words like "skateboard" and "ramp". Such a distance can serve as a proxy for semantic similarity and can be computed by measuring the distance between vectors in several ways, for example, cosine or euclidean distance. Plots of the waveforms and extracted features were produced with the LibROSA python package (https://librosa.github.io/librosa/) and Parselmouth (https://github.com/YannickJadoul/Parselmouth), a python library for the influential Praat software (Boersma and Weenink, 2018).

6.4 Introducing technology in psychiatry

Access to new technology is continuously transforming society in several ways, also healthcare, specifically in the management of psychiatric disorders. The use of technology for communication around health-related issues is of course nothing exclusive to the 21st century, and devices for tracking health information can be traced as far back as the middle ages when news about the bubonic plague was signalled across Europe by means of bonfires (Wootton, Craig, & Patterson, 2006). Recent years have seen a fairly substantial increase in the amount of information that can effectively be transferred using telecommunications channels, as well as a dramatic increase in usability via internet connected mobile devices. Healthcare in psychiatry has also benefited from this, where services via telecommunication have been accepted and even preferred. This was evidenced by the fact that mental health was the field where the use of technology for remote contact was employed most extensively (53% out of 383,565 visits) in a sample of 217,851 patients between the years of 2005-2017 in the US (Barnett, Souza & Mehrotra, 2018). This can be taken to mean that patients generally accept that conversations with clinicians can happen through digital channels. The importance of this cannot be understated, as the process of finding ways to make useful computations on verbal behavior, while not trivially easy, would be a lot more difficult if the patients were only willing to talk in face-to-face meetings.

However, to capture speech in a format that is available for quantitative analysis technological devices must be introduced into the clinical and research environment. Acceptance of this has changed over recent years, courtesy of the sheer number of mobile phones owned by an increasing percentage of the population. These devices have an array of sensors, including microphones, and are particularly suited for the deployment of small computer program applications, or “apps”, that can be tailored to the specific needs of research or healthcare services (Anthes, 2016). These programs can contribute with a multitude of different measurement modalities beyond recording speech, such as how much a person moves during the day, sleeps during the night or how they rate symptoms of stress or depression (e.g., Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015). The new possibilities for making measurements of behavior with digital devices has led to the proposal that this can lead to a new approach to phenotyping individuals (Insel, 2017), and has been promised to transform psychiatry into a more data-driven type of medicine (Hsin et al., 2019). In addition to the use

of specific applications it has also been argued that information gained from behavior on social media platforms should be utilized for research and mental healthcare delivery (Inkster, Stillwell, Kosinski, & Jones, 2016). Mobile applications have also been successfully employed in interventions for disorders such as depression (for a meta analysis: Firth et al., 2017), and can increase the value and precision of monitoring effect of mobile interventions in patients with schizophrenia (Schlosser, Campellone, Truong, Etter, Vergani, Komaiko, & Vinogradov, 2018; Bucci et al., 2018). Even if it is not at the core of the topic of this thesis, it is worth mentioning that there is emerging evidence that mobile device-based interventions can be effective for a wide range of disorders (Ebert et al, 2018; Linardon, Cuijpers, Carlbring, Messer, & Fuller-Tyszkiewicz, 2019), provided these can be developed and used in a safe and controlled manner.

Effective standards or guidelines for development of mobile applications for healthcare can have utility for implementation efforts in the future. Guidelines should concern aspects such information security and privacy protection, effectiveness of intervention, user experience and adherence optimization, and last but not least the possibility for integration of resulting data with other clinically relevant infrastructure such as medical records (Torous et al., 2019). While there are over 10000 mobile applications for mental or behavioral health at the time of writing (October - 2019), a standardized way of evaluation the quality of such apps has yet to emerge (Carlo, Hosseini, Renn, & Areán, 2019). Some notable efforts have surfaced from organizations such as the American Psychiatric Association (American Psychiatric Association, 2018) and the British Organization for the Review of Care and Health Applications (ORCHA, 2019). These efforts will be crucial to enable the clinician to select the most appropriate digital tools for their patients.

For any approach to be successful the tools will also have to be considered acceptable to use, namely enjoyable and appropriate for providing useful information for monitoring clinical states. Today users have high expectations from mobile tools because popular applications in daily use have excellent user experience, and there is an expectation of some value in return for time spent. Indeed, a survey by Yang, Maher, & Conroy (2015) found that only about one in four people are willing to spend more than 10 minutes per session with health related mobile applications, with only about two in four agreeing to spend more than 5 minutes. Considering the importance of usability engineering for successful implementation, it is

important to assess whether tools are considered useful and acceptable in the cohort of interest. In short, enjoyable tasks that are perceived to be useful will foster the collection of more and better data.

There have been a few successful implementations of specifically speech-based assessment methods using mobile devices (e.g., Faurholt-Jepsen et al., 2016). Research from our own group has also been able to leverage such acoustic parameters for modelling self-reports on affective states (Cohen et al., 2019a) and affect as perceived by external observations (Cheng et al., 2018). These are very promising results, but limiting analysis to acoustic measures (i.e., *how* something is said) may exclude important information about semantic content (i.e., *what* is said). Encrypted acoustic information about speech can be more comforting for the individual in terms of not running the risk of privacy violations of disclosure of what has actually been said in conversations, but there may be new possibilities from analyzing verbal behavior if the entire content of spoken utterances is available. Provided it is possible to find safe and legal solutions for collecting and moving raw speech data via digital channels, the utility of computational speech analysis would depend on the available technology for practically going about analyzing the collected speech samples.

6.5 Natural language processing for research and clinical utility in psychiatry

The use of natural language processing and semantic language models has already provided useful contributions in psychiatry (Corcoran, Benavides, & Cecchi, 2019). For example, Latent Semantic Analysis has been used to complement traditional ratings of incoherence in patients with schizophrenia, and in discriminating them from healthy control participants (Elvevåg, Foltz, Weinberger, & Goldberg, 2007; Iter, Yoon, & Jurafsky, 2018). Similar methods have also been applied to speech transcripts to discriminate patients with schizophrenia, first-degree relatives and healthy controls (Elvevåg, Foltz, Rosenstein, & DeLisi, 2010; Mota, Copelli, & Ribeiro, 2017; Mota, Sigman, Cecchi, Copelli, & Ribeiro, 2018), and also differentiating those at high risk for serious mental illness (because they are part of a family with high density of psychosis) from unrelated participants who seem healthy (Rosenstein, Foltz, DeLisi, & Elvevåg, 2015). It thus stands to reason that similar methods may be useful in studies that link language to genetics and indeed it has been shown that

language samples can be used to parse cognition in a candidate gene study (Nicodemus et al, 2014). Such methods have also been applied in vivo in neuroimaging and occurrences of off-topic incoherent speech been shown to be associated with lower neuronal activity in prefrontal areas of the brain. These findings directly connect objective measurements of coherence and incoherence in speech to the underlying neurobiology namely cortical activation (Hoffman, 2019).

Critically, the new methods do have clinical translation value as they can be used to discover clusters of pathological signs that are not elicited by or visible with traditional methods, and may predict onset of increased illness severity (e.g., psychosis) in groups or individuals at high risk (Bedi et al., 2015; Corcoran et al., 2018; Rezaii, Walker, & Wolff, 2019). Approaches that rely on detecting patterns of signs depend on accurate and timely measurements. If the new speech processing methods are going to have the translational value it is vital that the infrastructure for capturing speech data is solid and that the computational methods can in fact derive clinically relevant features from verbal behaviour in the real world.

6.6 The aims of the thesis

This thesis aims to address a set of questions confined to four topics, namely (i) infrastructure for data collection including speech, (ii) analysis of single spoken words, (iii) analysis of multiple-word spoken phrases and (iv) analysis of longer utterances of connected speech, all in the context of remote mental state assessment.

In Paper I, the fundamental infrastructure is examined that can enable the movement of mental state assessment via speech samples out of the controlled laboratory into real-word settings. Specifically, the following questions are addressed:

- Is it practically viable to capture speech data from a cohort of patients with psychiatric disorders using mobile devices?
- What are the key aspects for successful development and deployment of tools to remotely collect speech responses using mobile devices?

- Is a tool that is developed in the USA acceptable within a Norwegian (i.e., cross-cultural) setting?

These questions are answered by discussing anecdotes and lessons learned from development experience building mobile applications for mental state assessment in two different countries.

For the demonstration of computational analysis of spoken words three well-established neuropsychological tests were adapted. The three tests represent different perspectives on speech, such that results will aid in the description and evaluation of different pathological conditions.

In Paper II single-word utterances are explored to answer the following questions:

- Can remotely collected single-word utterances be automatically analyzed yielding results that are comparable to more traditional, labor-intensive methods in neuropsychological assessment?
- Can this method be used to measure attentional bias and control, and can it reveal differences between groups of patients and healthy volunteers?
- Can computational methods provide new ways of measuring speech response performance?

These questions were answered by analyzing speech responses from the Stroop color-word test. In this test, participants are asked to name the color of different words presented to them on a screen, and automatic speech recognition software was used to derive detailed characteristics of their responses. Response time latencies were used to measure inability to ignore salient properties of stimuli and infer possible biases in attention.

In Paper III multiple word utterances were examined to answer the following questions:

- Can the temporal aspects of the flow of speech when several words are put together in a sequence be quantified and visualized?
- Can the word-to-word semantic relationships and how they relate to speech production be quantified and visualized?

To achieve this spoken responses were analyzed from a simple, widely used and well tolerated task, namely a category fluency task. Participants were given one minute and asked to name as many animals as they could think of. This provided a small but fruitful model to explore speech production.

Paper IV moved beyond single words and assessed complete instances of connected speech, where participants were asked to retell stories they had been presented through the loudspeakers of a mobile device. Two questions were addressed:

- Can a computational procedure be derived for scoring the quality of a recall that is on par with human ratings?
- Is it viable to use automatic speech recognition in this kind of automated procedure, or will it introduce too many errors to be useful?

To answer these questions analysis software was constructed to compare the transcriptions of story retellings to the text of the original story that was presented. Performance of this analysis approach was compared between when the transcription of spoken responses was done by human transcribers and when words were transcribed by automatic speech recognition software. To be able to answer with affirmations to the two questions above, it was expected that there would be a high correlation between human ratings and machine scores, even if the transcriptions were done automatically.

7 Methods

7.1 The *d*MSE and the *MinTest* mobile applications

The setting for this exploration into computational analysis of speech was an international research project for mental state assessment tool development, supported by the Research Council of Norway (grant # 231395 awarded to Brita Elvevåg). An important goal for the project was to develop a practical tool and evaluate the feasibility of employing it for assessing the risk of harmful behavior in patients with severe mental illness.

Prior to developing the data collection tool, user needs were assessed in a survey format in clinicians (N=90) by examining what current practice is to psychiatric risk assessment and what clinicians considered might be ways to improve current methods (Cohen et al., 2019b). While there was high variability in the types of measures that clinicians endorsed for assessing risk, there were a number of commonalities in terms of general classes of assessment types used. Since these broadly fell into the categories of assessing cognition, motor skill and language, and 25 unique behavioral assessment tasks that assessed these domains were developed. These behavioral assessment tasks were integrated into a mobile application called the *delta* Mental Status Exam (*"delta"* to indicate our interest in *change* in mental states; *d*MSE) and eventually narrowed down to 12 separate tasks of various nature (Appendix 1 presents an overview of the tasks selected for the latest installment of the mobile application). The items were similar in form and structure to standardly employed neuropsychological tests, but were designed so that they could be remotely self-administered daily. The tasks were short and engaging and required listening, watching, speaking, and touching to interact with the smart device. A separate software version with the same tasks was developed for Norway and was named *MinTest*, which translates to "MyTest".

The apps were developed on the iOS platform (a mobile operating system created and developed by Apple Inc.). This platform permits fast development of highly usable interface components, easy speech recording and capture of touch screen responses. Additionally, this makes the testing procedures easy to deploy to smartphones (or lower cost internet-connected iPods), with the app available for download from the Apple software store.

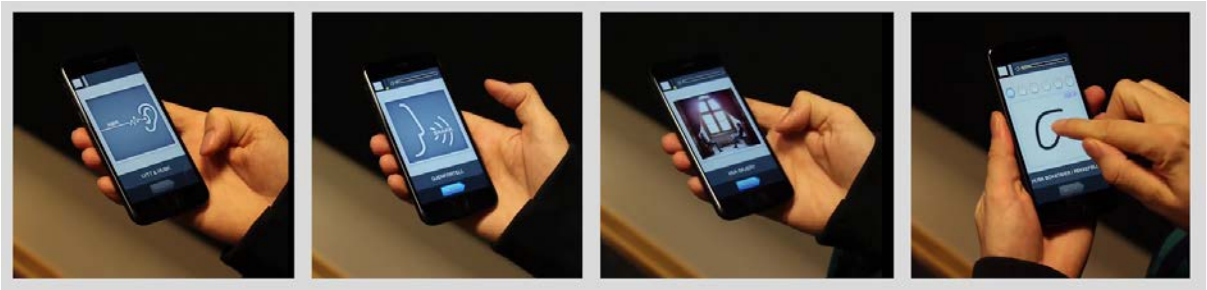


Figure 2.

The application required listening, speaking, watching and making touch-screen responses. These images are collected from the instructional video for the Norwegian mobile application, and the video can be viewed at https://vimeo.com/199976652.

7.2 Participants

A total of 353 participants contributed in three different studies over the course of 3 years (between 2014-2017). Two data collection phases were conducted in the US arm of the study, with the first phase using the full set of 25 assessment tasks, and the second phase using a narrower set of 12 assessment tasks. This approach with two phases of data collection in the US resulted in two rather different corpuses of responses that is drawn on for analysis in the exploration in this thesis, using subsamples of the participant cohorts that provided sufficient relevant data for the speech tasks that the individual analytic approaches focused on. In short, the participants who produced data for the individual investigations were sampled by convenience. Not all tasks were administered to all participants. This was an effort to balance the intention to test the feasibility of a wide variety of tasks while keeping the testing time to a minimum, also balanced with the intention to have a core set of tasks that could be used for longitudinal assessment. This means that the data presented in this thesis are not carefully controlled for possible differences between cohorts, and conclusions based on comparisons between groups of participants (e.g., patients versus healthy volunteers) should be done with care, given that the purpose of the overarching study was usability assessment rather than case-control comparisons *per se*.

The first phase recruited 25 patients with severe mental illness, with data collected in an outpatient setting, as well as 79 presumed healthy undergraduate students from Louisiana State University presumed to be healthy (henceforth “healthy volunteers”). Responses from all these patients were examined in Paper IV, where story recall performance from all

participants was analyzed. In the second phase of data collection, 105 patients at an inpatient facility for treatment of substance use disorders, as well as 120 undergraduate students were recruited. From this cohort, 57 patients and 86 students completed 5 sessions that presented the Stroop color-word test, and these participants were included in the investigation described in Paper II. Twenty-four patients and 35 students completed sessions that included a verbal fluency test, and responses from these participants were examined in Paper III. The discrepancy between the total number of participants included in the research program and the number of participants that was included in the investigations in papers II and III was partly related to the fact that not all assessment tasks were included in all consecutive sessions with the mobile application. For example, the task discussed in Paper III was only presented every sixth session. Summary statistics of key demographic variables of the cohorts can be found in Table 1.

The Norwegian cohort (N = 24, Table 1) consisted of inpatients recruited from a substance-use facility at the University Hospital of North-Norway, as well as stable outpatients at the same hospital that contacted the project after hearing about the mobile application under development. Healthcare professionals at the same hospital and otherwise interested healthy volunteers were also recruited for the assessment of acceptability and appropriateness of the application in a cross-cultural setting. The Norwegian participants were presented with a Norwegian version of the mobile application - *MinTest* - that corresponded to what was used in the second phase of data collection in the US.

Table 1 - Participants

Samples	N	% male	Age, Mean (SD)
Paper I: Norwegian usability study (N=24)			
Professionals	10	70	36.0 (10.9)
Patients	7	57	35.7 (8.9)
Healthy volunteers	7	43	35.0 (9.5)
Paper II: Substance use facility study (N = 141)			
Patients	57	100	39.1 (11.2)
Healthy volunteers	84	20	20.0 (1.9)
Paper III: Substance use facility study (N = 59)			
Patients	24	100	39.1 (10.7)
Healthy volunteers	35	12	19.5 (1.5)
Paper 4: Severe mental illness outpatients study (N=105)			
Patients	25	48	49.7 (10.4)
Healthy volunteers	79	62	21.7 (1.4)

7.3 Procedure and analysis

All participants were given instructions on how to use the mobile device and application prior to testing, and at a minimum, the first session was conducted in the presence of a research assistant. The first session also included a short standardized video that explained the procedures and demonstrated a selection of tasks. At the time of writing (October - 2019), both the US version of the video (URL: https://www.youtube.com/watch?v=TjseOrDf6BM), and the Norwegian version (URL: https://vimeo.com/199976652) can be found online, and can provide the reader with a clear illustration of the data collection procedure.

7.3.1 Paper I: Discussion of anecdotes from practical implementation

The main content of Paper I presents a perspective on the development of tools for remote mental state assessment, and as such was not a result of a specific experimental procedure. The issues of practical challenges are discussed though anecdotes from our development and implementation process, commenting on specific details on technical, legal and cultural issues.

To assess whether or not the Norwegian mobile application was acceptable to users, participants were to respond to a short online questionnaire before and after using *MinTest* (Appendix 2). The introductory questionnaire included questions on demographics and attitudes towards the use of smart devices in monitoring mental health. Attitudes were assessed by answering the question "In general, how do you feel about using mobile phones for monitoring mental health?" using an on-screen slider scale where extreme left was marked "Very against" (score = 0) and extreme right was marked "Very for" (score = 100). After the first session participants took the smart devices (iPhone SE or 5s) home to do the rest of the data collection remotely. When five or more sessions were completed, or the participant expressed that they wanted to conclude testing, there was a final meeting with the researcher for verbal and questionnaire feedback. The main acceptability outcome was whether or not users liked the application, responded to by answering the question "Did you enjoy using *MinTest*?" using an on-screen slider scale between "Did not like" (score = 0) and "Liked very much" (score = 100). In total, 152 testing sessions were collected in Norway, where results from three (2%) sessions had corrupted and unreadable data files. The average session duration was 12.2 minutes (SD = 2.2 minutes).

If technology has matured to the point that useful tools can be developed, as is claimed in this thesis, its would be expected that the discussions in Paper I could provide insights and key reference material to benefit researchers and clinicians who would want to leverage technology for the purposes of remote mental state assessment.

7.3.2 Paper II: Temporal measurements of single-word speech production in the Stroop color-word task

Each session with the mobile application contained one sequence of Stroop task trials, and participants were given verbal and written instructions before the session started (Figure 3). A visual prompt appeared before the sequence commenced, with the words: "SAY TEXT COLOR", and a vocal prompt saying "Say the color the word is printed in". The first word presentation was initiated by the press of touch-screen button from the user, then all subsequent presentations for the session appeared consecutively in a pseudo-random sequence for 96 seconds. In the trial, thirty-two words were presented in three stimulus conditions (8 congruent stimuli, 8 incongruent stimuli and 16 animal word stimuli. Congruent stimuli

consisted of color-words printed in the same color that they represent e.g. “RED” printed in the red color. Incongruent stimuli consisted of color-words printed in one of the remaining three colors (e.g. RED printed in green color). For baseline values of performance, animal words of 3-6 letters (DOG, BEAR, TIGER, MONKEY) were presented in all of the four colors. Words were presented on a white background in capital letters (Arial bold font, height = 165 pixels) and four different colors: RED, BLUE, GREEN and PURPLE (see Figure 3 for example of colors). Words stayed on the screen for 1500 ms, followed by a fixation cross for 1500 ms, resulting in a regular Inter-Stimulus Interval (ISI) of 3000 ms. The paradigm was based on the methods of Perlstein, Carter, Barch, & Baird (1998), with some notable adjustments, namely fewer trials and a shorter inter-stimulus interval.

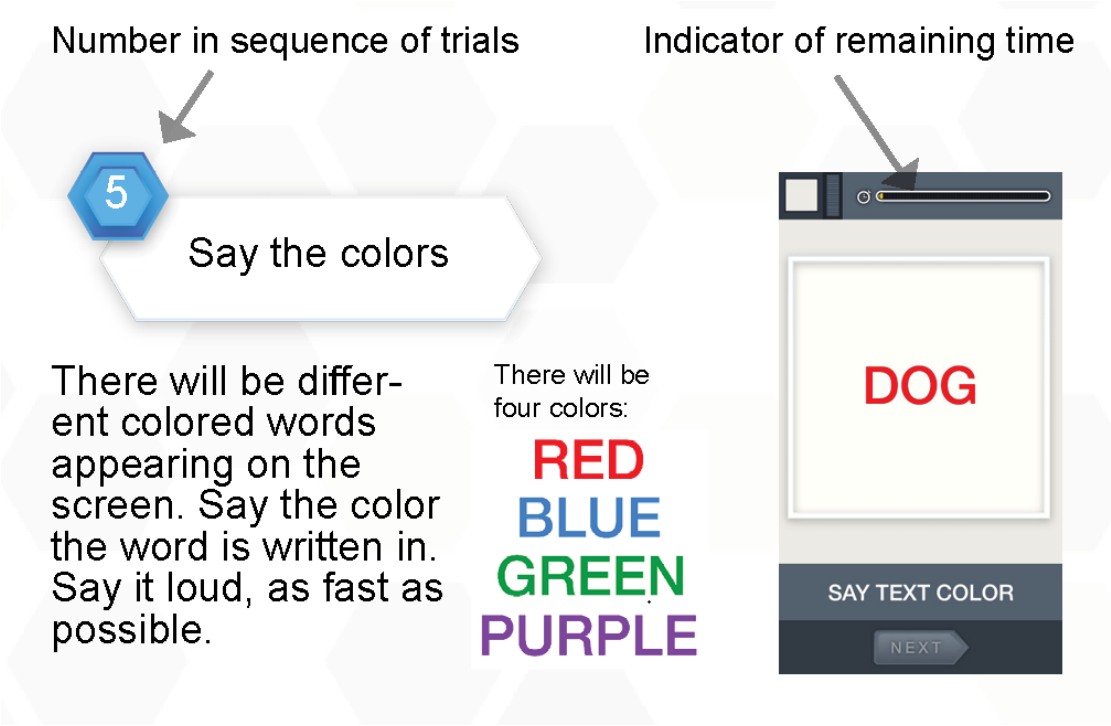


Figure 3.

An example screenshot from the written instructions for the dMSE (a more complete overview can be found in Appendix 1). For the Stroop task, the four possible colors to be named were displayed to reduce ambiguity about what would be considered a correct response word (e.g., to use “purple” not “violet”). The screenshot shows an example trial where the word “DOG” is presented in red ink, and the correct response would be to utter the word “red”, as fast as possible. The horizontal black bar at the top of the screenshot was filled with a yellow color from left to right as the trial progressed, representing how much time had passed and how much was left.

The spoken word responses were recorded continuously throughout the 96 seconds of the trial using the microphone built into the smart device, sampled at 16000 Hz and saved in a flac-format for further processing. The onsets of the speech responses (i.e., the moment when the

first sound waves of the response reached the microphone) were automatically timestamped at 10 milliseconds (ms) increments by an in-house developed automatic speech recognition model, using the Kaldi speech recognition toolkit (Povey et al., 2011). To increase the accuracy of the word recognition, the language model used was specifically tuned to recognize the relevant words in the Stroop task (i.e., the color words), taking advantage of the fact that words such as “Green” and “Red” were more likely to occur than “Car” or “Spoon”. The word error rate for the recognizer was estimated to be 6.26%.

The on-screen appearance of stimulus words was also time-stamped, and the response time latency (RT) was derived by calculating the duration between stimulus- and response timestamps. From the time-stamps general metrics of performance was derived. These were processing speed (average RT), intra-individual variability (coefficient of variation: Standard Deviation (SD) of RT divided by mean of RT, expressed as a percentage) and accuracy (as percentage correct responses). In addition, conflict-related metrics specifically assayed how conflicts between presented colors and word categories (i.e., congruent/animal-words/incongruent) affected the latency of responses. Word category conflict effects were calculated in two related measures: (i) *Interference*, expressed as the difference between the mean response time of the incongruent trials and the mean response time on purportedly neutral trials, and (ii) *Facilitation*, expressed as the difference between the mean response time of the congruent trials and the mean response time of the color-neutral animal-words trials. To expand beyond the traditional response time analysis, computation analysis of digital recordings also allow for measurements of acoustic properties of how a spoken response is uttered. For this analysis a simple feature was selected, response word duration, namely how long it takes from the start of an utterance to the subsequent silence after end of speech.

The statistical significance between participant groups and stimulus conditions (i.e., present or not) was assessed with repeated measures analysis of variance (rmANOVA) performed with the Analysis of Factorial EXperiments package, implemented in the R programming language (Singmann, Bolker, Westfall, & Aust, 2018). Simple group effect sizes were expressed with the Cohens’ *d* metric, and a fairly broad exploration using post-hoc Welch’s t-tests was conducted such that marginally significant differences should only be considered as suggestive.

Given that the procedure was based on a previously well-established paradigm and stimulus set (i.e., the selection of colors, the use of animal words) it was expected that group averages of response times should be somewhere in the range of 600 ms and 1200 ms, and that differences between conditions should be in the range of 50 ms to 200 ms (based on results from the original study; Perlstein et al., 1998). It was also expected that the incongruent, conflicting stimuli would result in longer delays, the interference hesitations. Patients were expected to have slower response times and group differences might interact with the stimulus type effects. If computational methods were comparable to traditional methods, these types of effects should be possible to measure also when conducting testing outside of the traditional laboratory setting. Additionally, exploratory analysis of the duration of the actual speech response utterances was conducted. Here the prediction was that healthy volunteers would produce faster, and more concise utterances as compared to the slower and maybe more slurred speech in patients, an assumption based upon the simple observation that psychiatric patients who are sick often talk slower, and this might be part of the illness or a side-effect of some neuroleptic medication.

7.3.3 Paper III: Timestamps and semantic vectors for multiple word sequences in a verbal fluency test

The goal was to chart the temporal dynamics of speech and objectively quantify the semantic relationship of response words in a category fluency task. This task was introduced in the second version of the mobile application, with the task appearing each sixth session (i.e., not all participants performed this task). Participants were given the following vocal prompt via the loudspeaker of the smart device: *"Name as many animals as you can. Any kind of animal, as many different animals as you can think of. You have up to 1 minute; start now"*. The trial started immediately after this prompt, and participants were given a visual cue on the device screen to start speaking plus a visual timer on the screen that indicated how much time remained for responding. These one minute long recordings of responses were collected via the microphone of the smart device. Recordings were transcribed in-house by humans and response-words timestamped using a forced temporal alignment procedure with the same type of ASR software as described for the methods in Paper I. The resulting word-strings were processed using regular string-editing procedures in the Python programming language.

Transcribed non-nouns, duplicates and words that were not in the category of animals (e.g., “Let’s”) were removed, and the remaining words were converted to their stem (e.g., “cats” to “cat”).

Semantic associations between successive word pairs were quantified using the word2vec cosine distance function implemented in the Gensim python package (Řehůřek & Sojka, 2010). The semantic space leveraged was a set of pre-trained, publicly available word vectors that had been calculated from approximately 42 billion tokens from the Common Crawl project with the GloVe unsupervised learning algorithm (Pennington et al., 2014).

It was predicted that participants would generate a rapid spurt of words followed by a general slowing such that the rate of word production (i.e., verbal fluency) that would decline gradually over the course of the 60s trial. In addition, it was expected that there would be a relationship between the semantic similarity and the delay of utterance between words. Within the sample, a higher verbal fluency was expected in healthy volunteers compared to the substance use disorder patients, as a result of multiple factors such as age, psychiatric comorbidity and medications.

7.3.4 Paper IV: Verbal memory recall of stories

Ten different text passages were developed and presented orally in a male voice via the mobile application. The participants were instructed to listen to the story and retell it with as many details as they could remember. Five of the passages were narrative stories and five of the texts were instructions on how to perform certain actions. The narrative stories were structurally similar to the Logical Memory subtest of the widely used Wechsler Memory Scale (Wechsler, 1997) and were between 69 and 87 words in length (average length = 75 words). Each narrative had two characters, a setting, an action that happened in the setting causing a problem, and then a resolution. The instructional passages were between 62 and 83 words in length (average length = 73 words), and started with a statement or question about an action that was to be performed, and then continued with a description on how to accomplish the goal of the action, and finally ended with some concluding details.

The participant was given a maximum of one minute to speak, and the time remaining was indicated by a timer bar on the screen of the smart device. This device recorded the participant's retelling and the recording ended after 60 seconds, or earlier if the participant concluded the trial by pressing an on-screen button marked "Next". All sessions with the mobile application included this task, and each participant was presented with one narrative and one instructional passage per session. They were additionally prompted to retell the narrative story later on in the testing session, by a prompt such as: "*Retell the balloon story again now. Put in all the details you can remember*."

Human raters within our team listened to the audio recordings of the recalls and assigned scores on a 0 to 6 scale, such that zero represented "silent or unintelligible", and a high score (6) indicated that all major and almost all minor concepts and themes were recalled. Every response was rated by between three and seven human raters and the average of these ratings was deemed the "gold standard" that the automated models aimed to predict.

Recordings were transcribed by both humans and automatic speech recognition. Human transcription was independently transcribed by two transcribers via an in-house designed web interface, and differences in transcription were then resolved, resulting in an overall human word error rate of 7.2%. Machine transcription was conducted on recordings that were pre-checked so as to not contain any personal private information. This was done using a fast, off-the-shelf service from Google (https://cloud.google.com/speech-to-text/), as well as with a custom speech recognizer based on the Kaldi speech recognition toolkit (Povey et al., 2011). Word error rates for the machine transcriptions were calculated by estimating the minimal edit distance with the Wagner-Fischer algorithm using the "jiwer" software package for Python (https://github.com/jitsi/asr-wer/).

The transcribed responses were preprocessed to provide text representations that were comparable in form to the original stories. This processing involved making all text lowercase and removing punctuation or non-speech related symbols from the transcription process (e.g., commas, periods or pound signs) as well as instances of transcribed hesitation markers such as "uh". This was done using the built-in string processing methods in the Python programming language (Python Software Foundation, https://www.python.org/).

Two quantitative measures of similarity between retelling and the original prompt was computed; a simple count of shared words and a semantic distance measure. The raw count consisted of the number of word types (i.e., individual words only counted once) that were in common between the two. The semantic space similarity measure was operationalized as the Word Mover's Distance between the recall and the original story, which is a metric of how far in semantic space one has to move all the words in the retelling to end up with an identical vector distribution as the original story (Kusner, Sun, Kolkin, & Weinberger, 2015). For this analysis, another set of publicly available pre-trained word embeddings than in the verbal fluency analysis was utilized, namely a semantic space with 300 dimensions derived from training a Word2vec model on 240 million words from the Google News corpus (Mikolov, Chen, Corrado, & Dean, 2013). This analysis was programmed using the Gensim software package (Řehůřek & Sojka, 2010).

The two similarity measurements, count of common words and the Word Mover's Distance between story and recall, were used as independent variables in an ordinary least squares regression model to derive estimated values of the human scores. The correlation between the estimated values and the average human rating was the main performance metric and was estimated using a 5-fold cross validation procedure. This procedure is performed to reduce bias in estimates, and involves dividing the data into 5 subsets, building the linear model on four of the subsets (i.e., the training sets) while leaving one subset (i.e., the test set) for estimating the correlation coefficient, and consecutively repeating the procedure such that all subsets had serves as both training and test sets. Both the linear models and the cross validation procedure were implemented using the scikit-learn Python module (Pedregosa et al., 2011).

Given previous results with similar methods and given the substantial difference between diagnostic groups, it was likely that healthy volunteers would see both a higher number of completed tasks, as well as higher recall ratings. If, as the main thesis states, computerized implementation and analysis methods were feasible then similar results when the testing of verbal memory was moved outside of the laboratory and into the hands of the patients themselves (i.e., self-administered)

8 Ethics

Since this research was conducted on smart devices outside of controlled environments, threats to the confidentiality, integrity and availability of research data were considered to pose important risks in the project, and important steps were taken to ensure the fundamental right to the privacy of research participants. Clinically relevant threats to the participants were also considered, such as the possibility that the data collection could exacerbate the severity of clinical conditions, but since traditional tests were employed (albeit in a digital format) the risks were deemed to be negligible and not further discussed here.

The author worked on the information security risk and vulnerability assessment for the Norwegian arm of the research project, and the report from this work can be provided upon request. The report outlined the main threats to participants and institutions that were relevant to the project, specified the likelihood of the threats occurring, and summarised this into a matrix where the severity of risks were classified (Figure 4). Stated briefly, the most severe risks were related to human errors in data management, and risk mitigation strategies were also described in the report.

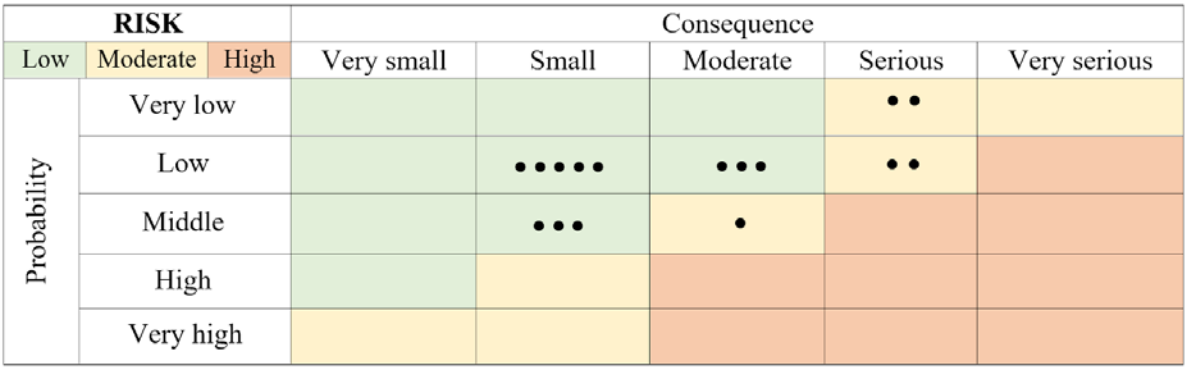


Figure 4.

An adapted version of the risk matrix from the Norwegian arm of the research project. Each dot represents an unwanted event category, which was identified by a specific code in the risk assessment report proper for further discussion. For example, the author identified five low-risk unwanted event categories that had a low probability of occurring, and a small potential consequence for the participants and institutions. The event categories with "Moderate" risks were carefully analyzed and mitigation strategies described.

The US part of the study was approved by the Louisiana State University Institutional Review Board (#3618) and all methods were performed in accordance with the relevant regulations and guidelines. To be included, participants had to be able to legally offer informed consent,

choose to offer written informed consent (see consent form in Appendix 3), watch an instructional video highlighting the risks, rewards and expectations of participation, and demonstrate an understanding of the study by passing a quiz with questions about the details of the study. Students were rewarded with course credits for participation, while patients were given monetary rewards of $5 per completed testing session with the application. Likewise, in Norway, the study was approved by the Regional Committee for medical and health research ethics for North Norway (#2014-85). Participants were provided oral and written information about the study and signed paper consent forms (i.e., not digital, see the consent form in Appendix 4) prior to completing an introductory questionnaire and then performed their first *MinTest* session in the presence of the investigator. Participants in Norway were all included in a lottery for this study where the winner received an Apple iPad Mini2.

9 Results

The main finding from this investigation was that it was indeed possible to derive digital data from the spoken words of patients with a variety of psychiatric disorders using mobile technology, and that this data could be analyzed with a variety of approaches to complement and expand on traditional mental state assessment (Holmlund et al., 2019a). In the following these results are reported by providing answers to the specific questions posed in the introduction.

9.1 Paper I: Moving assessment out of the lab - Infrastructure for data capture

Is it practically viable to capture speech data from a cohort of patients with psychiatric disorders using mobile devices?

Yes, it was possible to remotely collect high-quality speech recordings from 353 research participants, including 137 patients with various psychiatric disorders. This means that new technological frameworks can provide unprecedented opportunities for self-administered behavioral and clinical assessments, where it is possible to participate in easy-to-use digital versions of traditional speech assessment tests as well as new variants that are suitable for use on a daily basis. This proof-of-concept study allow us to confidently move forward with improving methods based on the lessons learned from this implementation.

What are some of the key lessons for successful development and deployment?

The study showed that solid usability engineering is crucial to ensure utility and acceptability with users and produce data that is comparable to more traditional testing methods. This can be illustrated by how classic tasks from neuropsychological assessment like the Stroop color-word Test (Stroop, 1935) can be converted into a format that allows for high frequency testing (e.g., daily) for longitudinal testing approaches spanning weeks or months.

Moving psychological assessment out of the laboratory setting (Figure 5, panel A) results in a number of technical challenges. Employing such remote data capture methods, both locally and internationally, necessitates that the technological infrastructure is sufficiently secure so

as to ensure the confidentiality and integrity of data transfers. This was solved in two different ways in this project, manually using cables and storage devices (used in the Norwegian setting; Figure 5, panel B) and via automatic transfers over the internet (used in the US setting; Figure 5, panel C). 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 within the countries involved that regulate such transfers. These same legal frameworks that regulate transfer and storage of personal data also grant participants strong rights, and participants have the right to request deletion of their own data at any point. For the researcher, this means that the frameworks require development of quite a sophisticated data management infrastructure. Tracking down and deleting all entries when an individual participant requests it can be challenging in many cases, the data management methods of the future need to be able to do this efficiently and safely by employing dynamic consent procedures and attaching privacy policies to data where possible. The fact that raw speech recordings are personally identifiable in their very nature, combined with the fact that participants can reveal highly sensitive information about themselves (or others) when prompted to speak, demands that researchers should default to adhering to the strictest legal frameworks and the safest technological solutions.

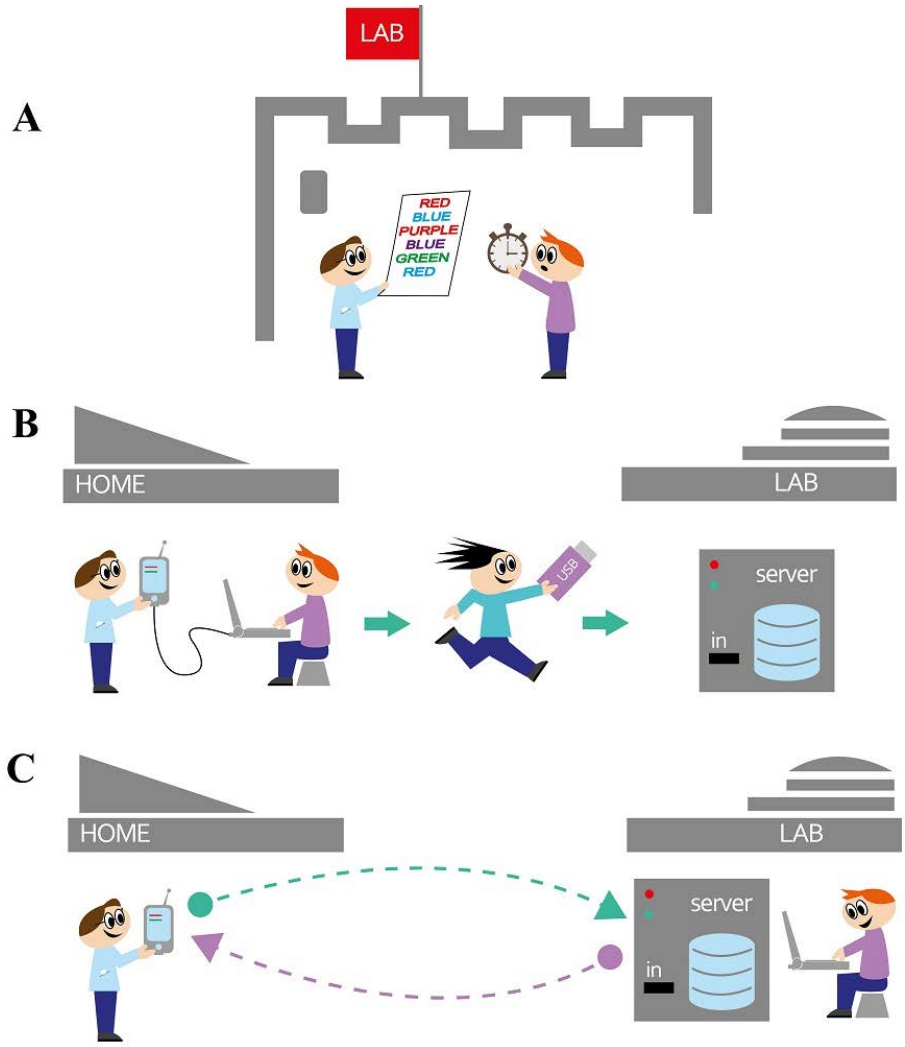


Figure 5.

Panel A: Traditionally, psychological assessment has been conducted in safe laboratory setting. Panel B: Digitalization of tests can enable testing outside of the laboratory, but moving the data from the hands of the individual back to the laboratory for analysis takes careful planning, and can be done manually with labor-intensive and error-prone procedures using cables and USB-drives. Panel C: Using online transfers are much more effective, but must be compliant with complex country-specific regulations.

Assessing cognition via the medium of language requires a careful consideration of how different cultures can affect results. For large-scale implementations, tasks need to be suitably translated and normed within the various languages and cultures. 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 expected to observations regarding how language is used as well as how contextual factors may affect behavior. Nonetheless, in designing these tools to assay psychological functions, and ultimately the integrity of an

individual's cortical function, great effort should be expended to ensure the tasks are language-neutral and culture-fair.

Was a tool that was developed in the US acceptable within a Norwegian (i.e., cross-cultural) setting?

Yes, participants appeared to accept the application, as evidenced by the amount of actual data that was captured in Norway, as well as from the questionnaire feedback given completion of the study. Participants gave an average acceptability score (i.e., "*Did you like using MinTest?*", 0-100) of 77.0 (SD 16.3). Complaints were mostly related to the session durations were too long (33% of users; see Figure 6 Panel A). Complaints about duration was not considered to be related to culture specifically, as preferences for shorter duration sessions has also been reported in US cohorts in other studies (Yang, Maher, & Conroy, 2015). While participants were asked to use *MinTest* for five sessions, some enjoyed it to the degree that they continued to use it for over a month (mean number of sessions completed was 7; median was 5, with a range between 1-42 sessions).

It should be noted that the small sample of Norwegian participants self-reported positive attitudes towards using mobile phones to monitor mental health, both before launching the *MinTest* app the first time (mean score on the 0-100 slider scale was 85.0, SD 16.8) and similarly at the end of their participation (mean score of 85.4, SD 13.6).

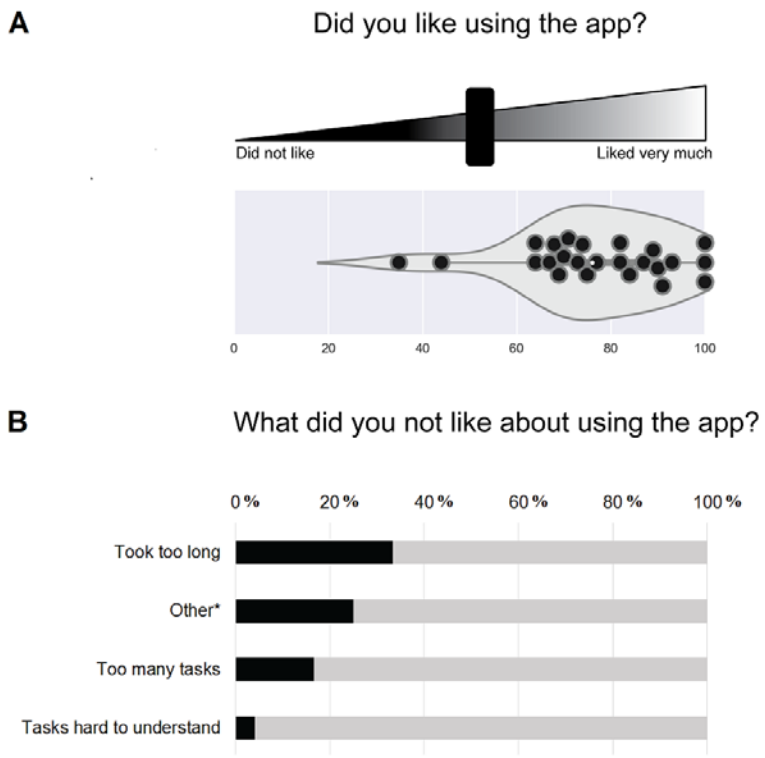


Figure 6.

*The cross-cultural acceptability of the US-developed mobile application was assessed in 24 Norwegian participants. Panel A: To measure acceptability, participants were asked to respond to the question "Did you enjoy using MinTest?" by placing the slider between "Liked very much" (Score = 100) and "Did not like" (Score = 0). The average score was 77.0, with only two responses below 50. Panel B: Black horizontal bars represent the percentage of participants who marked tick-boxes with the respective reasons for not liking the app. *Statements could be entered in free text: Comments on the combination of duration and number of tasks, tasks too difficult to perform, lack of tutorial, lack of variation in tasks and one comment on dislike of own frustration when performing poorly. This kind of detailed feedback is crucial to further development of the mobile application.*

Appropriateness was measured by the extent to which the participant believed the tool could be useful, and the question *"Do you think MinTest be useful for monitoring mental health?"* was answered on a slider scale between "Not useful" (score = 0) and "Very useful" (score = 100). The mean appropriateness score was 76.5 (SD 15.1), findings taken to indicate that participants considered the tool to be potentially useful.

In sum, for small teams of researchers who aim to deploy more complex spoken language assessment approaches the technical, legal and cultural challenges sum up to a situation where the development of such tools may not seem feasible (Figure 7). The solution to this will be increasing multidisciplinarity in teams with methodological, technical, legal and cultural

expertise to ensure high levels of user and clinician satisfaction along with superior data quality and safety.

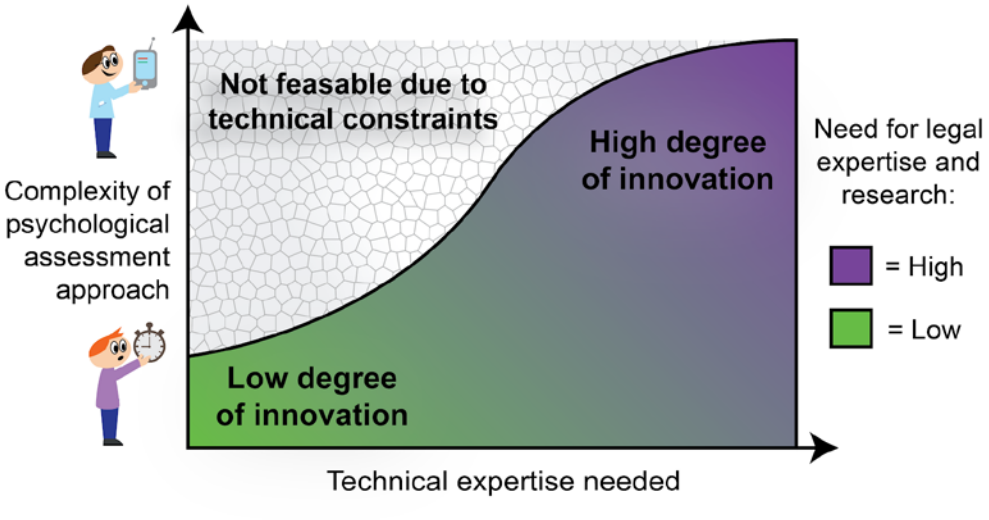


Figure 7.

The relationship between complexity of the digitized psychological assessment tool, the technical expertise needed to use it properly, and the resulting legal implications. Performing multimodal remote assessment can prove not feasible if there is little technical expertise available (gray area). Using pen-and-paper methods can be safe, but with a lower potential for innovation (green area). Highly technical approaches can be innovative, but demand expertise in complex legal domains (purple area). New legal domains may emerge and warrant legal research in their own rights, as technological advances move research- and clinical practices beyond the scope of current legislature.

9.2 Paper II: Single word speech production in the Stroop task

Can remotely collected single-word responses be automatically analyzed yielding results that are comparable to more traditional, labor-intensive methods in neuropsychological assessment?

Yes, clear differences were observed between the different stimulus word conditions, namely that the actual color or animal words interfered with and delayed response times for the naming of the ink colors. This was represented by an extremely robust main effect of stimulus condition in a repeated measures analysis of variance ($F(2,282) = 522.4$, $p < 0.001$). Additionally, there were dynamic relationships in responses that could be leveraged to describe performance over time. Response times became faster with practice (main effect of

practice session: $F(4, 524) = 26.7$, $p < 0.001$), but the differences between conditions, the interference effects, remained over the five testing sessions (Figure 8). The procedures was also able to detect what words were uttered, and the percentage of correct responses was different between conditions ($F(2,282) = 40.5$, $p < 0.001$).

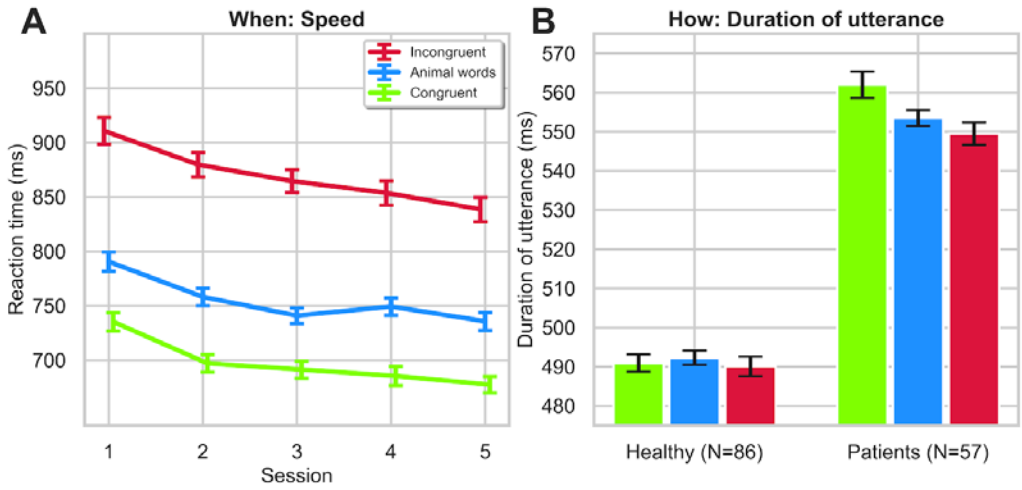


Figure 8.

Results from the Stroop color and word task. Panel A: The delay between an on-screen appearance of the written word and the speech response depended on word category conditions. Response times were longest when the ink colors were incongruent with the meaning of the words (e.g. "GREEN" written in red ink; red line in plot) and shortest when the word and ink were congruent. The animal word stimuli resulted in intermediate delays. Participants became faster with practice, represented by the sloping lines between the daily sessions. The presented data are combined across both healthy volunteers and patients, and error bars represent the standard error of the mean for the response times. Panel B: The duration of the word utterances was shorter in healthy participants as compared to patients, indicating a faster way of expressing the words, not only that the delay, or "hesitation", after stimulus presentation was different between groups. Patient responses on congruent trials were also of longer duration compared to responses on incongruent conflict trials.

Can this method be used to measure attentional bias and control, and can it detect differences between groups of patients and healthy volunteers?

Probably yes. It was possible to reliably measure stimulus condition effects on the order of 50-100 ms, which should be sufficient for detecting differences between response patterns in the ability to ignore highly salient words. The groups differed in the delays (i.e., hesitations) when naming colors of animal words versus of congruent color-words (Cohen's $d = 0.63$, $t = 3.53$, $p = 0.003$). This is encouraging because it points towards a possibility to use other word categories, for example words related to intoxicating substances, making a self-administered and spoken version that can be highly suited for adaptive and tailored testing purposes. More generally, the assays were sensitive to differences between patients and healthy volunteers in

both overall response time speed (Cohen's d = 0.71, t = 4.01, p < 0.001) and variability (Cohen's d = 0.69, t = 3.93, p < 0.001).

Can computational methods provide new ways of measuring speech performance?
Yes. As a proof-of-concept of what speech processing technology can add to the analysis of spoken responses the duration of response utterances was clearly different between healthy participants and patients (Cohen's d = 1.01, t = 5.9, p < 0.001). This promises to be a useful metric and a radically new approach analyzing single word responses in this paradigm. In addition, it was possible to detect stimulus condition effects on utterance duration (main effect of condition: F(2,282) = 4.1, p = 0.018), but these effects were small compared to the temporal resolution of the measurements.

In sum, the study successfully demonstrated in this study how a combination of two mature tools, namely automatic speech recognition and the Stroop test, could be leveraged in new ways to provide descriptions of fine-grained details of the temporal dynamics in verbal expressions. The methods were possible to automate and thus easier to implement in large scale studies and longitudinal monitoring compared to manual methods.

9.3 Paper III: Speech production and semantic coherence in the verbal fluency task

Can the temporal aspects of the flow when several words are put together in a sequence be quantified and visualized?
Yes, timestamps of words supplied by speech recognition software allow for easy plotting and quantification of word production. First of all, counting the number of words per minute was performed automatically, and it was possible to show differences between patients and healthy volunteer participants on the traditional word count measure. Healthy participants generated 25 words (range = 8-36) and patients generated on average slightly fewer (mean = 19 words; range = 11-36; $t(57) = -4.0$; $p < 0.001$).

The speed of speech production was possible to illustrate by demonstrating trajectories on a graph that plotted word counts versus time. Figure 9, panel A, is an example of such a graph,

where a steep section of the plot represents an epoch of fast speech. In paper III two noteworthy examples were chosen, namely a healthy person who generated 31 words versus a patient who generated only 12 words in the one minute period. In addition illustrated how much variability there can be in individual data as compared to group means (Figure 1, panel B in Paper III).

We did not venture into statistical analysis of different temporal epochs (e.g., faster production in the first part of the trial versus last part of the trial), as such findings are rather uncontroversial and have been documented previously (e.g., word-production in consecutive 60s windows in a 3 minute task (Elvevåg, Weinstock, Akil, Kleinman, & Goldberg, 2001); and word-production in 15 s windows in a 1 minute task (Krukow, Harciarek, Morylowska-Topolska, Karakuła-Juchnowicz, & Jonak, 2017)). Looking at the data from this thesis with such methods is certainly possible, as demonstrated in Figure 9, panel B, where delay between words in 15 s windows is used as a metric of speech production (e.g., 0-15 s, 15-30 s and so on).

Can the word-to-word semantic relationships and how they relate to speech production be quantified and visualized?

Yes, by using word embedding methods it was possible to quantify the cosine distance between successive words, a function of whether there is a high level of similarity in word pairs or not. The word to word similarity was demonstrated to fluctuate considerably over the one minute duration of a trial. Semantic similarity between successive words was negatively associated with the speed of speech, meaning that overall there were longer pauses between successive words when they were semantically unrelated. This was represented by a significant negative correlation between inter-word delays and the cosine distance between them ($r = -0.36$, $p < 0.001$).

It was further explored whether the word sequences were more semantically similar at the beginning of the trial, as compared to the end (Figure 9, panel C). This exploration was not presented in Paper III, as it was intended to be a short communication regarding new technology, and could potentially be a good demonstration of how temporal data can be leveraged in combination with word embedding in novel ways to better understand the

dynamical nature of speech. Overall, semantic similarity changed over the course of the trial (Figure 9, panel C). Earlier time windows demonstrating higher semantic similarity between the response words (i.e., words that are more commonly used together), confirmed by a significant main effect of Time Window ($F(3,141) = 28.4$, $p < 0.001$) on a 2*4 repeated measures ANOVA. Critically though, there was no significant main effect of Group ($F(1,47) = 0.1$, $p = 0.753$) or Time window*Group interaction ($F(3,141) = 2.2$, $p = 0.090$).

It should be noted that these types of analyses with temporal windows have clear limitations, notably because the choices of window size lacks a strong theoretical foundation and are magnitudes slower than the 10 ms temporal resolution in current speech recognition methods. More to the point, these images of the raw data presented in Paper III should inspire other researchers to think of new ways to use these techniques to more effectively probe the relationship between verbal fluency and mental states.

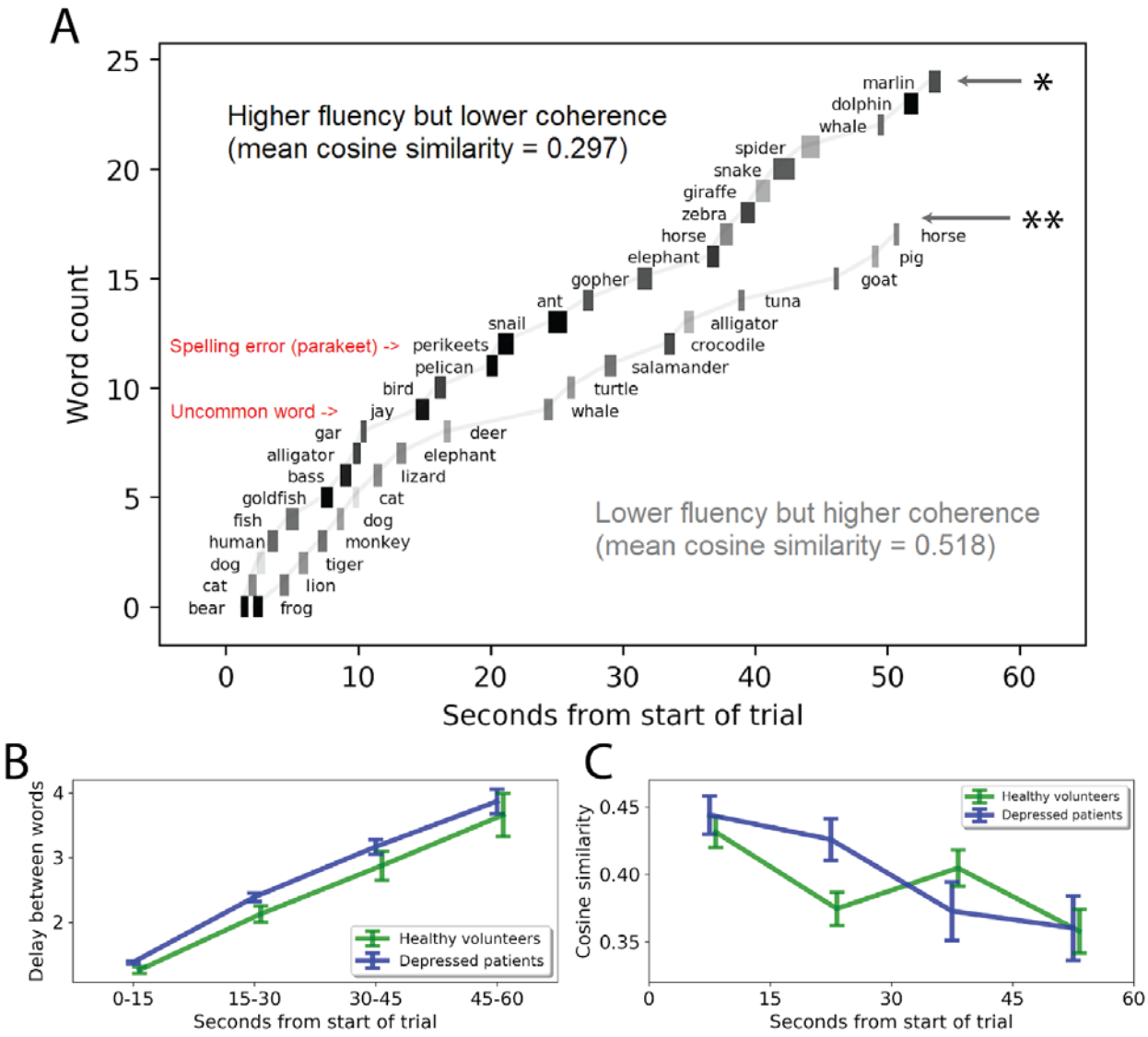


Figure 9.

*Graphical representation of both temporal and semantic information that can be extracted from a verbal fluency test. Panel A: Two example trials where one shows higher fluency (24 words per minute, marked *) compared to the other (17 words per minute, marked **). The utterances can be read from left to right. For each identified word the location on the y-axis increases, creating a presentation where a steep slope represents higher fluency at that point in time. Word-to-word similarity is represented in grey-scale, such that words that have a lighter box have lower distance to the previous word in a semantic vector space. The width of the box represent utterance duration, such that words that were pronounced over a longer time interval have wider boxes. Panel B: The delay between words increased overall during the course of the trial, equally for both groups. Panel C: The semantic similarity between word decreased over the course of the trial. Error bars represent standard errors of the means.*

9.4 Paper IV: Computational analysis of recall performance in the verbal memory task

Can a computational procedure be derived for scoring the quality of a recall that is on par with human ratings?

Yes, it was possible for speech recordings to be converted to suitable text data using automatic speech recognition. This data could then be analyzed using natural language processing methods to derive a score of the amount of story details remembered.

First, it was important to establish that it was possible to collect useful recall data remotely. Ninety two percent of a total 1035 collected speech responses were amenable to further processing (86% for patients). The retellings were on average 61 words (healthy participants' mean = 62 words, SD = 21, and patients' mean = 49 words, SD = 22; Cohen's d = -0.8, t = -9.1, $p < 0.001$), with more short responses in patients (for example "I don't remember"; Responses under 10 words: healthy = 5%, patients 20%).

Second, ground truth scores were established by averaging the ratings from human raters. Ratings from each of the 3-7 raters correlated with the average rating at R = 0.83 (ranging between R = 0.73-0.89), and it was this level of reliability of rating that a robust automated procedure would be expected to operate within. As expected, the average rating was higher for responses by healthy participants (Mean = 4.6, SD = 1.1) than by patients (Mean = 3.3, SD = 1.3, Cohen's d = -1.1, t = -9.1, $p < 0.001$).

Third, a simple count of the words that were in common between the transcriptions and the original story was highly predictive of human ratings, with the count correlation to the average human ratings at R = 0.82. The Word Mover's Distance between the recall and the original story correlated with the average human raters (R = - 0.81) and combining both the word-count and distance measure in an ordinary least squares linear regression model could predict human ratings at a level of performance on par with individual human raters (R = 0.83, range 0.74-0.90 across 5 cross-validation folds). The notion that the automated procedure provided valid scores was strengthened by a finding that computed ratings for

retellings from healthy participants were higher (Mean = 4.6, SD = 0.9) as compared to those from patients (Mean = 3.4, SD = 0.9, Cohen's d = 1.3, t = 15.4, $p < 0.001$).

Is it viable to use automatic speech recognition in this kind of automated procedure, or will it introduce too many errors to be useful?

Scores based on automatic speech recognition transcription correlated well with human ratings and this was considered a viable approach. Error rates in machine transcriptions were high, particularly for patient data, and should promote caution when interpreting results.

Using an off-the-shelf speech recognition service the overall word error rate was 23.3% (43.7% in patients), but even so, the predictions of a combined feature model based on these transcriptions correlated with human ratings at R = 0.80 (range 0.74 - 0.88 across five folds).

The word error rate using the customized ASR system was notably lower, with an overall word error rate of 10.5%. (24.8% on speech from patients). Computed ratings based on these transcriptions correlated with the average of the human raters at R = 0.82 (range 0.74 - 0.88 across five folds), which was in the range of human to human agreement of 0.73 to 0.89.

Despite high word error rates, the predicted ratings from fully automated procedures correlated highly with results derived using the procedure where humans transcribed the recordings (R = 0.96-0.99). This robustness can be attributed to the fact that errors are common on non-essential words (e.g., “is”, “and”, etc.), with the more important content words generally being transcribed correctly.

In sum, a fully automated pipeline with remote presentation of stories, response detection and automatic rating of recall quality seemed to be a feasible future solution. High error rates for automatic transcriptions will be an important challenge, but should be amenable to improvements using language models customized to specific tasks, similar to those employed in Paper I. Encouragingly, the complex type of behavior analyzed there does not represent simple, overlearned responses like naming colors in the Stroop task or listing animals in the fluency task, but can provide us with an opportunity to effectively measure memory performance, the ability to recall events.

10 Discussion

This project successfully collected and computationally analysed speech data from patients with psychiatric disorders as well as healthy volunteer participants. Through this implementation effort valuable lessons were learned about the importance of interdisciplinary collaboration where clinical, technical, and legal expertise is needed to ensure safe and effective speech analysis platforms. Automatic speech recognition could be utilized to derive valuable metrics of exactly when a word was spoken, if the right word was used, and it could characterize the way words were uttered. Furthermore, these techniques could provide descriptions of word-to-word flow when multiple words were spoken, revealing temporal patterns that also showed a theoretically sensible correspondence to the semantic content of the words. Lastly, longer utterances of speech where participants recalled stories they had heard were analyzed, resulting in computed scores of recall performance. In sum, the findings support the claim that technology has matured to a point where it is possible to derive objective observations of signs of disease from assays in impaired speech production or decline in memory function.

Our measurements were able to capture the crucial connection between the semantic aspects of language and the temporal aspects of speech. In the Stroop task there was a conflict in meaning between the colors and the printed words that was transferred into observable perturbations in speech behavior, namely measurable differences in delays before responses. The crucial connection was also evident in the verbal fluency task, where words that were less semantically similar had longer delays between them. Therefore the argument is made in this thesis that the strong link between semantic and temporal aspects strengthen a notion that research into language in psychiatry should employ vocalized responses and avoid paradigms that require button-presses or self-report scales. Another significant shift in thinking in terms of traditional psychometrics is the possibility of getting good quality speech data for assessment purposes outside of the laboratory and that this data is generated by the persons themselves on a regular basis. On a more somber note, the promise that precision medicine naturally follows from increased ability to collect data is not quite true, as it is now necessary to conduct a significant amount of research to develop ways to create useful models of the dynamic and often incomplete longitudinal data from individuals.

The progress in technology that has been made available to us would not be possible without recent advances in statistical modelling and computer hardware. In addition to the increased access to mobile devices, there has been development of more complex methods for using artificial neural networks for computation (i.e., "deep learning"; LeCun, Bengio, & Hinton, 2015), where speech recognition in particular has been seen as an area that benefited greatly from the advances (Hinton et al., 2012). In the case of this thesis, the deep learning methods are employed as tools for small parts of the analysis presented in this thesis, connecting word labels to sound pressure waves or creating the word vector embeddings needed to calculate semantic distances. Exactly how these methods can produce such excellent performance can be opaque to most readers, leaving researchers and clinicians with a sense of mystery that can erode the trustworthiness of the methods. Indeed, we as researchers should strive to build and implement explainable, transparent and generalizable models for psychiatry (Chandler, Foltz, & Elvevåg, in press). By combining this progress in computational methods with the large body of knowledge in the medical domain, a very useful convergence of human and machine intelligence may emerge, where the term "artificial intelligence" is more synonymous with "computer aided" (Topol, 2019, 151). As this progress continues, it will be possible to employ these methods on different levels of analysis, potentially resulting in robust, reliable and generalizable pattern detecting tools for capturing pathological behavior.

In addition to speech analysis, the *d*MSE and *MinTest* data collection platforms were also able to supply valuable information from patients regarding their subjective experience of affective states. The findings from these symptom-based investigations were not the focus of this thesis, but are worth mentioning to provide context and demonstrate a wider scope of what is possible to gain by deploying digital tools such as the *delta* Mental State Examination and *MinTest*. The investigations explored self-reported feelings of hostility (Cowan et al., 2018), aggressive urges in schizotypy (Le et al., 2018) and loneliness (Le et al, 2019). These domains are crucial for gaining access to data that are relevant to an important goal of the research project, namely creating a tool for assessing risk of harmful behavior in patients with psychiatric disorders. A key finding from these publications was that to explain clinically meaningful variables such as self-reported hostility towards others it was not enough to build models from static demographic data, but it was crucial to take into account *more dynamic measures* of mental states such as concurrent self report on negative affect and acoustic parameters in speech (Cowan et al., 2018). In addition to the traditional focus on validity and

reliability, the “resolution” of measurement tools will be important in detecting the relevant signals of disease. Improving both temporal resolution (i.e., frequent measurements), spatial resolution (i.e., accurately defining the context of investigations) and spectral resolution (i.e., having multiple types of measurements e.g., self-report combined with acoustic metrics in speech) will be important for future advances in mental state assessment (Cohen et al, in press). What remains to be discovered is how these lessons can be turned into actual clinical tools, providing healthcare workers with actionable information based on multiple streams of data (Holmlund et al. 2019b).

10.1 Limitations

Given how the content of this thesis spans many domains, from discussions of cultural issues down to detailed technical aspects of analytical procedures, a wide variety of limitations can be discussed. Therefore, a selection of the most pressing issues are approached in this section even if it is acknowledged that other issues remain unaddressed. First addressed is the issue how the very strict requirements to privacy and information security has the drawback of hindering progress in the field. Then the generalizability of the experimental results are discussed, particularly in terms of understanding the psychiatric disorders and neurocognitive functions that were studied. Lastly there is a discussion about whether or not repeated measurements of function necessarily will lead to a better description of mental states.

While great care in protecting the right to privacy is called for in Paper I, there is a risk that overly stringent demands for security aspects create significant hurdles for progress. It is tempting to be overly careful, with a painstakingly slow approach, but this does not provide practical solutions. Arguably, some of the legal frameworks dependent on what is classified as identifiable information rely on an outdated legal thinking, as it is currently possible to use many different sources of data to re-identify individuals (Al-Azizy, Millard, Symeonidis, O’Hara, & Shadbolt, 2016). In fact, too much restriction on implementation efforts will lead to very few researchers and clinicians experienced with emerging techniques, and where the actual use-cases that are captured can bias the knowledge of usability aspects greatly. As an example, the number of participants in the Norwegian usability study was low and may have had an overrepresentation of individuals who were more positive towards technology and more capable of enjoying the *MinTest* tool. If the security requirements for exploratory

implementations like this had been lower (i.e., allowed for internet-transfer of data) data larger and more representative sample of the population would have been collected. Nonetheless, the sheer number of participants was well within an estimated number of participants that should be able to provide an appropriate cost/benefit ratio for detecting important issues, also within the different participant categories (Nielsen & Landauer, 1993). A more sinister issue can come to mind if one considers the retardant effect of excessive security requirements, namely a cost for patients who can not benefit from a future technology with real time feedback, potentially ending in a catastrophic event that could be avoided if caretakers were provided with information in a timely manner. Ultimately, evaluations of safety depends on us being cognizant of the relationship between consequences and likelihood of adverse events and being explicit about what are considered to be acceptable risks in research and clinical settings, similar to the risk and vulnerability assessment in this project (mentioned on page 43, Chapter 4).

A main limitation with the results presented in this thesis lies in the generalizability of the findings to inform on the disorders that are discussed, notably psychosis and schizophrenia. While investigating specific differences between patients with psychiatric disorders and healthy individuals was not the core purpose of the study, such comparisons was included to illustrate how the measurements were able to differentiate the different groups. For example, the 25 patients in the verbal memory assessment experiment in Paper IV were considerably older than the healthy volunteer participants. This confounding factor alone could explain the large differences in both the ability to complete tasks on a mobile phone, courtesy of practice with such devices, and recall performance proper, courtesy of age-related decline in cognitive functions. In the same study, the sample size of 25 patient participants may be considered low, particularly in light of the diagnostic heterogeneity of the sample (e.g., schizophrenia, bipolar disorder, etc.). However, this issue of small samples can be said to apply to several other recent studies that do case-control comparisons for natural language processing measures in psychiatry (e.g., Rezaii, Walker, & Wolff, 2019) reporting classification test characteristics with a validation set of only 10 participants. This is not to say that the emerging results in the field of clinical computational language analysis are necessarily false, but it is important to be cognizant of possible overfitting of models to the small but precious patient samples that are available for research. Even if the groups are larger and it is possible to incrementally “control” or “adjust” for important metrics such as age and gender with

traditional methods (e.g., multiple regression analysis), the risk of identifying nonexistent associations are high (Westfall & Yarkoni, 2016). Having said that, future studies that seek to leverage the power of machine learning approaches for psychiatry will need to be on a much larger scale, resulting in an impossible prospect to control variables in the traditional clinical sense (e.g., strict exclusion criteria). This may actually be less of a problem as the large variability will in fact be an advantage in the subsequent characterizations, courtesy of the very manner in which machine learning works. The challenges in participant sampling severely limits the utility of the traditional approach of case versus control group comparisons in this field. These methodological challenges underlines the importance of having effective longitudinal designs where relative change in performance can be properly quantified in a larger, more varied sample of patients and healthy individuals.

In the case of the Stroop experiment in Paper II, deconstructing the performance measurements in a manner to usefully reveal definitive and specific information about attention is difficult. Specifically, the paradigm does not offer a truly semantically neutral stimulus condition, and it is likely that the animal words used in addition to the color words may have produced particular interference effects in their own rights. Since the main goal of this study was not parsing these functions per se, but rather it was to replicate the classic effects in a new setting, this issue was not catastrophic in terms of the utility of the results. Likewise in terms of generalizability, the healthy participants were different from the patients in several important ways, perhaps most notable was the fact that the patients were being treated by a wide variety of medications that could affect the function of the central nervous system. Medication effects might also have been present for some of the presumed healthy participants, as they were not specifically screened for medication use. Unlike traditional experiments, control parameters such as the visual angles of Stroop stimuli (for example, eyes 86 cm from the screen, letter size corresponding to an exact number of degrees of the visual field, etc.) were not controlled since the task was presented on small mobile screens at arbitrary lengths from the eyes of participants. This was considered less of a problem since the classic effects were replicated, and the replication is important because it indicates that certain strict constraints on task administration are not crucial for the effects to be found in robust paradigms like the Stroop task.

For the quantification of semantic coherence in the word-to-word relationships of Paper III and the story-to-recall similarity in papers IV a host of other approaches could be considered. For example, the word-to-word coherence measure is highly local one could explore different window sizes (e.g., words #1-3 versus word #4-6, etc.). Such a sliding window approach could reveal more long-ranging patterns. The number of features explored in the model for predicting human ratings of recall in Paper IV was purposely modest and could be expanded upon. In particular, one should seek features that are not collinear. Linguistic measures with a more narrow definition such as increased usage of ambiguous pronoun usage (Iter, Yoon, & Jurafsky, 2018), reduced usage of possessive pronouns (Corcoran et al., 2018) and referential anomalies in noun-phrases (Çokal et al., 2018) could lead to a more holistic description of the utterances. More broadly, an important discussion considers whether the use of vector space models and measures of distance in semantic space is sufficiently specific for defining language incoherence. Consider the two words "car" and "boat". Given that they are both means of transportation they may not be able to catch the absurdity if a patient said the sentence: "*I travelled across Arizona in my boat, as per usual*." Such issues may be improved by a more specific selection of corpuses and more sophisticated methods to build the language models, but there is still the possibility to lose some of the nuances of the descriptions developed from decades of clinical research and practice (e.g., Andreasen, 1986).

Beyond the limitations of the more specific neuropsychological tests it is also possible to question the assumption that increased frequency of repeated measurements will lead to a better understanding of the underlying pathology. Cognitive functions, when assayed by applications using traditional tests such as the Digit Span, Trail-Making test and Stroop test may provide useful biomarkers of neurodevelopmental problems in patients with schizophrenia (Melle, 2019), but high temporal resolution of longitudinal follow-up may not be necessary, as the changes happen in adolescence and there is no clear evidence of cognitive decline from psychosis prodrome to first-episode psychosis (Carrión et al., 2018). Therefore, even if it is stated several times in this thesis, is it really worth making repeated measurements? The answer will of course depend on the specific functions that are assessed, for example one can perhaps expect higher utility for this for understanding highly state-depended conditions like panic disorder compared to more trait-dependent conditions such as severe autism spectrum disorder. Obviously it is difficult to conclude at this point in time since the very tools described here will have to materialize in the clinic and be implemented at

scale. Even taking these challenges to usefulness into account, there are indications that measures of intra-individual (i.e., within-person) variability in cognitive performance are valuable in their own right (MacDonald, Li, Bäckman, 2009). Importantly, the manner in which performance fluctuates from day to day should inform more individualized operationalizations of what constitutes meaningful or significant change in performance over time, for example a change expressed as within-person standard deviations (Salthouse, 2007). Going even further, there is evidence that intra-individual variability in behavior has important links to brain structures and neuronal activity and may provide warnings of underlying pathology (MacDonald, Nyberg & Bäckman, 2006).

10.2 What will these new technological approaches mean for psychiatry?

The findings presented in this thesis have implications that are both positive, in terms of what psychiatry research can achieve in the future, but also challenging, in terms of the demands that will be put on the infrastructure and even new types of specialist personnel in clinical care (e.g., Noel, Carpenter-Song, Acquilano, Torous, & Drake, 2019). The methodological building blocks described here can create robust and tailored instruments, but the remaining and persistent challenges in terms of privacy protection must be faced with infrastructure that stands on a solid footing. This solid footing has not yet materialized for psychiatric applications. This is the main relevance of Paper I of this thesis, namely to identify the importance of interdisciplinary collaboration for technological infrastructure. Previously, scientific progress could be made by having a room with two chairs, a well-crafted interview guide and paper and pencils for note-taking. While traditional interviews can still be useful tools in many respects, a research and clinical infrastructure that includes microphones for speech data collection, mobile devices for multiple other measures, internet based transfer of information and timely feedback about data to clinicians and patients alike will be needed for true translational value.

The opportunities for clinical utility of voice analysis that are closest on the horizon are likely to be tools that can assist in classifying the observed mental states as belonging to well-known categories such as “mania”, “depression” or “psychosis”. This is due to the current focus on

machine learning approaches, where cases of pre-labeled instances of a phenomenon can be used to "train" algorithms to identify similar patterns in new, unseen data. Faurholt-Jepsen and colleagues (2016) have achieved impressive results by examining the acoustic features of speech from telephone calls made by patients with bipolar disorder. They found that it was possible to classify the clinical states of mania (Area under Receiver operating curve; AUC = 0.89) and depression (AUC = 0.78) with acceptable accuracy. The difference in classification performance, namely that methods are better at identifying mania, has also been demonstrated by others (Karam et al., 2014) and can provide clues as to what conditions are most likely to benefit from voice-based approaches for detecting disordered mental states. More broad categorizations are also possible. In our own research program it has been found that participants could be classified as belonging to either "healthy" or "patient" groups with a sensitivity of 0.80 and a specificity of 0.74 (F1-score = 0.76), based only on speech samples from a verbal memory task (Chandler et al., 2019). It is important to note that any effort to use classification results based on these new methods for clinical purposes must be done with great care such that they do not lead to overdiagnosis and overtreatment (Vogt, Green, Ekstrøm, & Brodersen, 2019).

Entirely new categories of clinically relevant phenomena may be possible to capture using remotely collected speech data. While the Stroop task traditionally has been employed to probe for disordered cognitive control, the amount of information that can be extracted from very small amounts of speech elicited with innocuous tasks may illustrate the new possibilities for assessing mental states. Words that evoke strong feelings in individuals can produce a larger degree of hesitation (e.g., spider-related words can in patients with spider phobias induce interference effects on the order of 190ms; Watts, McKenna, Sharrock, & Trezise, 1986), potentially providing a roadmap towards finding signals of what the individual speaker considers important or salient. While it is certainly a stretch of imagination given the current state of assessment methods, one could envisage methodology that could analyze utterances and detect the occurrence of aberrant assignment of salience to certain environmental events (for a discussion of "aberrant salience", see Kapur, 2003). A tool for detecting aberrant salience can be very helpful for identifying development of psychosis in individuals. Clues of aberrant salience may also come from analysis based on acoustical parameters from the same data, with a combination of features translating into effective tools for psychiatry.

Aiming to provide solutions for detecting crucial conditions such as psychosis or mania demands implementation infrastructure that can provide reliable information for clinicians in a timely manner. Returning to the methods described by Faurholt-Jepsen and colleagues (2016), the oscillations in clinical state of bipolar disorder demands frequent measurements and a strength of their system is a well-developed mobile platform that can allow for a bi-directional feedback loop between patients and health care providers (Faurholt-Jepsen et al., 2014). Their platform is also capable of collection more than speech data, as they have access to on-device sensors such as accelerometers for activity tracking, and can collect self-reports of mental states by various means. It is when voice data can be combined with other data types (e.g., activity tracking) that it is possible to achieve what we have described as increased spectral resolution (Cohen et al., in press). In short, approaching the phenomenon of disordered mental states from multiple angles at once will result in a more clinically valuable picture. If methods can address one of the ultimate challenges, namely to be accepted by the patients and clinicians, there is a good chance that they will be able to reliably provide real-time feedback from quick and automatic analysis thus potentially being life-saving for patients whose worsening clinical state could have consequences such as self-harm or harm to others.

In addition to providing practical tools for monitoring mental states, new data collection platforms such as the *d*MSE and *MinTest* will improve models of human behavior. Similar to how the example from the introduction assumed that a model of a giraffe with appropriate detail on anatomical structures can give us a better idea of the appearance and function of a real giraffe, the findings regarding the temporal dynamics of a verbal fluency task gave us appropriate description of details on how the verbal fluency tapered off over the course of a minute. With more widespread implementation of methods that can make descriptions of such dynamics, the expected “shape” of verbal fluency over one minute can better inform assumptions about verbal behavior. Several important assumptions in psychological science been challenged in recent years due to the failure to replicate many of the studies that much of the field is founded upon (Aarts et al., 2015). This is an issue that obviously has deep and multifaceted roots, and may be part of the reason why some will go as far as saying that a field of interdisciplinary research under the umbrella “cognitive science” has yet to fully materialize (Núñez, Allen, Gao, Miller Rigoli, Relaford-Doyle, & Semenuks, 2019). The

methods of data collection described in this thesis may be a part of the way forward towards a more coherent field of research where data are not collected in isolated labs. Although the idea of moving forward with better ecological measurements has been conceptualized for many years, psychiatry research is now poised to make a significant change to how psychiatric science is conducted since the technology that is needed for data collection have become more available as off-the-shelf services. In the continuous oscillation between data-driven and theory-driven progress, it is very likely that the current influx of ecologically valid data will lead to better scientific models of mental states.

This discussion will end by returning to where the introduction started, namely on the relationship between symptoms and signs. Teasing apart and understanding the different concepts of symptoms, signs and other behavioral aspects related to a putative underlying physical pathology can have a significant impact on how best to provide treatment (e.g., see Waddel, Bircher, Finlayson, & Main (1984) for an interesting perspective on pain-related behavior). Being able to objectively track verbal output from a patient could, and should, be an important part of the apparatus that determines treatment plans for the use of psychotropic medications. Even today in the 21st century in psychiatry, research findings and new instruments at all levels (e.g., genes, networks, behavior) are validated against gold standard measurements based on self report of symptoms (e.g., the Structured Clinical Interview for DSM; First, Spitzer, Gibbon, & Williams, 2002). The circularity and disadvantage of this cannot be understated as this problem permeates the entire field and stagnates progress. Harsh as this may sound, there is nonetheless a growing consensus that the old way of thinking about symptoms is out of date (e.g., the National Institute of Mental Health Research Domain Criteria; Insel et al., 2010). This thesis is in keeping with this newer non-phenomenological approach to mental health problems, namely that a detailed operationalization and measurement of the actual signs in patients is fundamentally necessary. The methods described in this thesis allow for precise measurements of *when* something is said, *what* is said, *how* it is expressed and should therefore provide parameters useful for tracking signs of treatment effects. When this tracking additionally can be done by leveraging telecommunication infrastructures regardless of location, it is possible to create a solid set of mobile tools to monitor mental states.

11 Conclusion

It is possible to obtain high-quality speech data from tasks administered remotely using mobile devices, and this data can effectively be analyzed to extract a wide variety of features that are relevant to mental state assessment. These methods of computational analysis of speech are not limited to the use of mobile devices. As long as there are microphones available for data capture, and as long as the involved persons have consented to the processing of speech, there are vast opportunities for effective, precise and valid assessments of mental states. Implementing this type of technology is not trivial and there is a need for specialized researchers and engineers to work together to address the several technical, legal and cultural challenges.

12 Final remarks and future perspectives

Moving forward, it is important to make measurements available in a *timely* manner, make them more *understandable* in terms of individual and *relative* change, and ultimately that results that can fit into a larger framework of the natural sciences.

Analysis must be conducted and results provided to clinicians and users within a relatively short time-frame to be useful. The investigations described in this thesis have a large disadvantage in terms of deployment in the clinic, namely that the time it has taken from the words were spoken by the research participants, to the time that relevant analysis was ready on computer screens has been on the order of years. For true translation value of the methods, results must be provided magnitudes faster, namely within minutes or hours. This will require safe, transparent and carefully crafted data pipelines that can translate from speech to on-screen information in a matter of seconds or minutes, rather than years. In a forthcoming book chapter on language assessment is psychosis this has been emphasized, and a hypothetical system has been described that can provide useful information to the clinicians desktop in a manner of seconds (Holmlund, Fedechko, Elvevåg, & Cohen, in press). Such a tool, capturing speech and operationalizing verbal behavior via the computational semantic methods the actual words expressed, and the manner of their expression through acoustic parameters, can provide the necessary ingredients for a test suite for cortical function and its dysfunction

(Figure 10). Akin to the way results from analysis of blood samples are presented, so too in the case of speech samples, presentations of data in formats familiar to clinicians (e.g., tables and graphs) may increase adoption. Such a presentation of data is possible in the near future and will be able to provide useful second opinions about speech expressions to future clinicians.

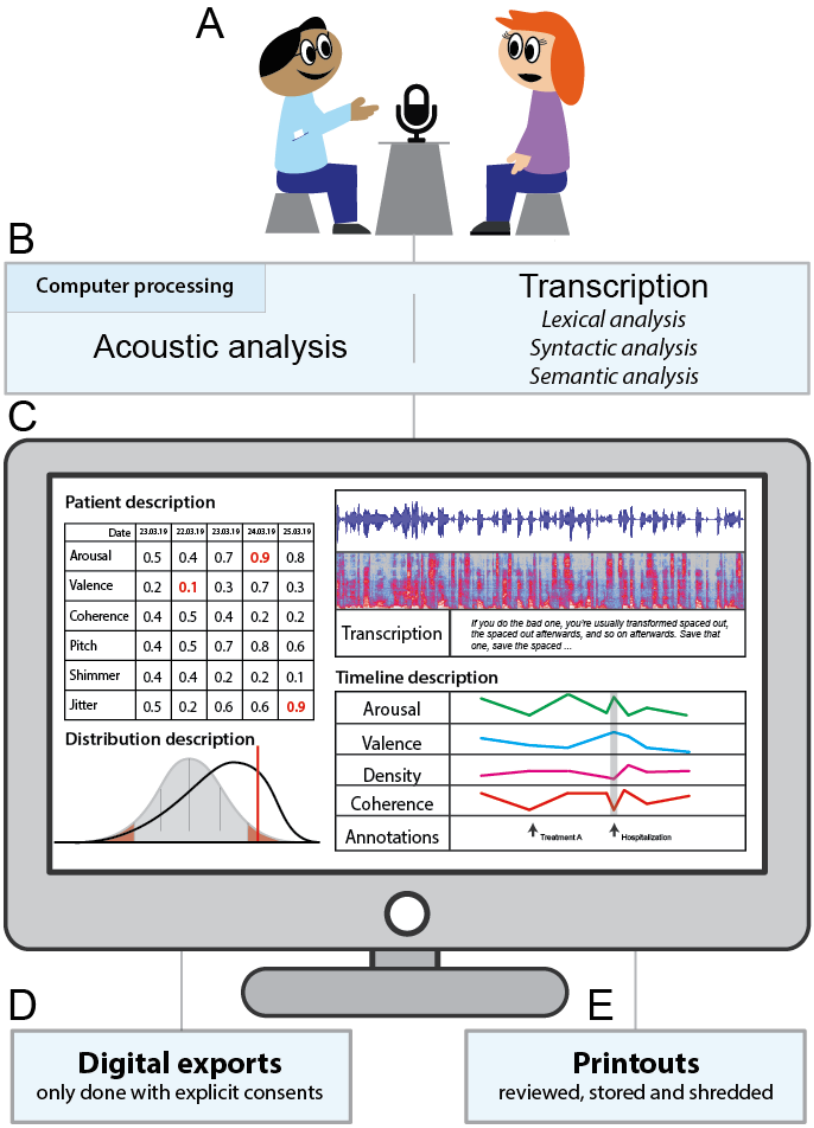


Figure 10.

An outline of necessary components for moving computational analysis of spoken words into clinical practice. Panel A: A major challenge to adoption will be introducing microphones in hospital settings. The microphones should be hidden and no covert processing should take place as it would be catastrophic to the patient's trust towards the system. Panel B: For full utilization of the methods described in this thesis, a computational device must be able to process both acoustic features and transcribe speech for natural language processing. Panel C: Adoption from clinicians will depend heavily on effective presentation of data. Key figures and trends should be presented in a format that is familiar but still allow for a new insights into the verbal behavior of patients by expressing how they relate to population and individual norms. Panel D: If data from recordings are to be exported outside of the immediate setting there must be explicit consent for this from the patient. Panel E: A possible first step for making speech analysis results available to clinicians could be to implement local, air-gapped systems that do analysis, display and prints of results. This may seem like a very old-fashioned way of

dealing with data export, but it is one that is familiar to clinicians and in fact still very common in hospital settings (e.g., electrocardiograms, echocardiograms). The local or "edge" processing has the added benefits of lower complexity in terms of information security (e.g., papers can be shredded).

In addition to translating speech to actual results faster, streams of multiple data types must be combined appropriately into a coherent analysis framework. This thesis has described how to derive results from tasks separately (i.e., the Stroop task, verbal fluency, verbal memory), but greater value can be gained when in the future data from tasks can be combined into multidimensional representations of performance in a single session of testing. Knowing how to integrate the separate scores into putative meaningful signs of disordered mental states is difficult. This should come as no surprise, afterall modern neuropsychology has benefited from a century of data collection by hundreds of institutions in order to build up databases that afford "norms", comparisons with other patients with similar pathologies, similar ages and gender. Thus it stands to reason that even in today's data rich digital society it will take many years to coordinate consortiums to safely share data for such purposes, even if the progress in data collection methodology continues at the current pace.

A wide array of analytic approaches can become feasible in a framework where speech data is continuously generated, but the fact remains that measurements must be presented in a way that is understandable to those who need them. One challenge lies in determining what to be considered "abnormal", and to represent the data in a way that is grasped quickly for decision support. One wildly successful way of presenting what can be considered normal development over time is growth-charts, as can be used for follow-up of pregnancies or infants (Cole, 2012; World Health Organization, 1995). Employing a normative function approach can help avoid the problematic concept of an "average patient" (Marquand et al., 2019). Successful applications has been seen in certain cohorts such as Attention Deficit Hyperactivity Disorder (Wolfers et al., 2019), and the approach has recently proved valuable in mapping the phenotypes of patients with schizophrenia and bipolar disorder (Wolfers et al., 2018). Acquiring this kind of detailed and longitudinal samples from individuals can also base assessment on the shape of the distribution of measurements of the individual person. The output of such assessments in the form of measurements plotted on a scale with clear indication of median values, interquartile and interdecile ranges are familiar to many clinicians, courtesy of the ubiquity of growth charts. For example, measurements that are outside of the 10th percentile are easily identifiable as unusual and may warrant attention.

Such charts may seem like overly simplistic proxies of the multidimensional nature of the data available, but can take the same dimensionality into account and potentially serve as a gentle introduction for many to how technology can provide behavioral signs of psychiatric illness.

Moving beyond effective presentation of collected data, an important path forward will be optimizing descriptions of speech for incorporation in a larger framework of medicine and into future models for predicting behavior and mental states. To build a solid foundation for describing behavior in psychiatry the field should strive to employ units from the International System of Units (SI; Bureau International des Poids et Mesures, 2019). For example, descriptions of psychomotor activity in patients that are expressed as acceleration (meters/seconds^2) and energy expenditure (Joules/minutes/kilograms; Faurholt-Jepsen, Brage, Vinberg, Christensen, Knorr, Jensen, & Kessing, 2012) are likely to be more useful and will be a better fit for integrating results into a larger framework with mature tools for dynamical modelling and time-series predictions (e.g., AutoRegressive Integrated Moving Average models; Ward, 2002). How speech and language can be described most adequately to fit within such frameworks with SI units is still unclear. A possible path forward may be presented by looking to the field of information theory (Shannon, 1948), where communication can be described in the units of “bits” and entropy. Some interesting progress has been made in describing the amount of choice related to individual words, and therefore opens up the possibility of measuring the degree of disorder, or entropy, of language expressions. These findings seem to relate to general traits of the human species, as patterns in the entropy differences between words (Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho, 2017) and the rate of information transferred per second of speech (Coupé, Oh, Dediu, & Pellegrino, 2019) across several languages around the world. A possibility of creating descriptions of behavior expressed in bits and joules is encouraging for the prospect of unifying research progress in psychiatry with the rest of the natural sciences.

13 References

Al-Azizy D., Millard D., Symeonidis I., O'Hara K., & Shadbolt N. (2016) A Literature Survey and Classifications on Data Deanonymisation. In: Lambrinoudakis C., Gabillon A.(eds) *Risks and Security of Internet and Systems. CRiSIS 2015. Lecture Notes in Computer Science*, vol 9572. Springer, Cham. doi: 10.1007/978-3-319-31811-0_3

Aleman, A., Hijman, R., de Haan, E.H.F., & Kahn R.S. (1999) Memory impairment in schizophrenia: a meta-analysis. *American Journal of Psychiatry, 156,* 1358–1366. doi: 10.1176/ajp.156.9.1358

Allampati, S., Duarte-Rojo, A., Thacker, L. R. Patidar, K. R., White, M. B., Klair, J. S., . . . Bajaj, J. S. (2015). Diagnosis of Minimal Hepatic Encephalopathy Using Stroop EncephalApp: A Multicenter US-Based, Norm-Based Study. *The American Journal of Gastroenterology, 111*(1), 78-86.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA.

American Psychiatric Association (2018). App evaluation model. https://www.psychiatry.org/psychiatrists/practice/mental-health-apps/app-evaluation-model

Andreasen, N. C. (1982). Should the term "thought disorder" be revised? *Comprehensive Psychiatry, 23*(4), 291-299. doi: 10.1016/0010-440X(82)90079-7

Andreasen, N. C. (1986). Scale for the Assessment of Thought, Language, and Communication (TLC). *Schizophrenia Bulletin, 12*(3), 473-482. doi: 10.1093/schbul/12.3.473

Andreasen N. C., & Grove, W. M. (1986). Thought, Language, and Communication in Schizophrenia: Diagnosis and Prognosis. *Schizophrenia Bulletin, 12*(3), 348-359. doi: 10.1093/schbul/12.3.348

Anthes, E. (2016). Mental health: There's an app for that. *Nature, 532*(7597), 20-3. doi: 10.1038/532020a

Aarts, A., Anderson, J., Attridge, C., Attwood, P., Axt, A., Babel, J., ..., Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943-943. doi: 10.1126/science.aac4716

Barnett M. L., Ray K. N., Souza J., & Mehrotra A. (2018). Trends in Telemedicine Use in a Large Commercially Insured Population, 2005-2017. *JAMA, 20,* 2147–2149. doi: 10.1001/jama.2018.12354

Bedi, G., Carrillo, F., Cecchi, G., Slezak, D., Sigman, M., Mota, N., . . . Corcoran, C. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia, 1*(1), 15030. doi: 10.1038/npjschz.2015.30

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., Kandola, J., Hofmann, T., Poggio, T., & Shawe-Taylor, J. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*(6), 1137-1155. doi: 10.1162/153244303322533223

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-I-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy, 19*(6), 275. doi: 10.3390/e19060275

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal, 38*(3), 218-226. doi: 10.1037/prj0000130

Bleuler, E. (1911). Dementia praecox oder Gruppe der Schizophrenien, 1st ed. Diskord, Leipzig, Wien.

Boersma, P., & Weenink, D. J. M., (2018). Praat: doing phonetics by computer (Version 6.0.37) [Computer program]. Amsterdam: Institute of Phonetic Sciences of the University of Amsterdam.

Bokat, C. E., Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: a meta-analysis. *Schizophrenia Research, 164*, 73-78. doi: 10.1016/S0920-9964(02)00282-7

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624-652. doi: 10.1037/0033-295X.108.3.624

Bousfield, W. A., & Sedgewick, H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30*, 149-165. doi: 10.1080/00221309.1944.10544467

Bousfield, W. A., Sedgewick, H. W., & Cohen, B. H. (1954). Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology, 67*, 111-118. URL: https://www.jstor.org/stable/pdf/1418075.pdf

Bucci, S., Barrowclough, C., Ainsworth, J., Machin, M., Morris, R., Berry, K., . . . Haddock, G. (2018). Actissist: Proof-of-Concept Trial of a Theory-Driven Digital Intervention

for Psychosis. *Schizophrenia Bulletin, 44*(5), 1070-1080. http://dx.doi.org/10.1093/schbul/sby032

Bureau International des Poids et Mesures (2019). SI Brochure: The International System of Units (SI). URL: https://www.bipm.org/en/publications/si-brochure/

Carlo, A. D., Hosseini G. R., Renn, B. N., & Areán, P. A. (2019). By the numbers: ratings and utilization of behavioral health mobile applications. *npj Digital Medicine, 2*(1). doi: 10.1038/s41746-019-0129-6

Carrión, R. E., Walder, D. J., Auther, A. M., Mclaughlin, D., Zyla, H. O., Adelsheim, S. . . . Cornblatt, B. A. (2018). From the psychosis prodrome to the first-episode of psychosis: No evidence of a cognitive decline. *Journal of Psychiatric Research, 96,* 231-238. doi: 10.1016/j.jpsychires.2017.10.014

Chandler, C., Foltz, P.W., Cheng, J., Bernstein, J.C., Rosenfeld, E.P., Cohen, A.S., **Holmlund, T. B.,** & Elvevåg, B. (2019). Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., Loveys, K. (Eds), *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Minneapolis, Minnesota, USA, June (pp. 137–147). URL: https://www.aclweb.org/anthology/W19-3016

Chandler, C., Foltz, P. W. & Elvevåg, B. (in press). Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness. *Schizophrenia Bulletin.* doi:10.1093/schbul/sbz105

Cheng, J., Bernstein, J., Rosenfeld, E., Foltz, P. W., Cohen, A.S., **Holmlund, T. B.** & Elvevåg, B. (2018). Modeling Self-Reported and Observed Affect from Speech. In: *Proceedings Interspeech*, Hyderabad, India, 2-6 September (pp 3653-3657). doi: 10.21437/Interspeech.2018-2222

Cirillo, M. A., & Seidman L. J. (2003). Verbal declarative memory dysfunction in schizophrenia: from clinical assessment to genetics and brain mechanisms. *Neuropsychology Review, 13*, 43-77.

Cohen, A. S., Fedechko, T. L., Schwartz, E. K., Le, T. P., Foltz, P. W., Bernstein, J., Cheng, J., **Holmlund, T.B.** & Elvevåg, B. (2019a). Ambulatory vocal acoustics, temporal dynamics and serious mental illness. *Journal of Abnormal Psychology, 128*, 97-105. doi: 10.1037/abn0000397

Cohen, A., Schwartz, E., Le, T., Foltz, P., Bernstein, J., Cheng, J., . . . Elvevåg, B. (2019b). Psychiatric Risk Assessment from the Clinician's Perspective: Lessons for the Future. *Community Mental Health Journal*, 1-8. doi: 10.1007/s10597-019-00411-x

Cohen A. S., Schwartz, E., Le, T., Cowan, T., Cox, C., Tucker, R., Foltz, P., **Holmlund, T. B.**, Elvevåg, B. (in press). Validating digital phenotyping technologies for clinical use: the critical importance of "resolution." *World Psychiatry.*

Çokal, D., Sevilla, G., Jones, W. S., Zimmerer, V., Deamer, F., Douglas, M., . . . Hinzen, W. (2018). The language profile of formal thought disorder. *npj Schizophrenia, 4*(1), 18. doi:10.1038/s41537-018-0061-9

Cole, T. (2012). The development of growth references and growth charts. *Annals of Human Biology, 39*(5), 382-394. doi: 10.3109/03014460.2012.694475

Corcoran, C. M., Benavides, C., & Cecchi, G. (2019). Natural Language Processing: Opportunities and Challenges for Patients, Providers, and Hospital Systems. *Psychiatric Annals, 49*(5), 202-208. doi: 10.3928/00485713-20190411-01

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, ... , Cecchi, G. A. (2018) . Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67-75. doi: 10.1002/wps.20491

Coupé, C., Oh, Y., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances, 5*(9), eaaw2594. doi:10.1126/sciadv.aaw2594

Cowan, T., Le, T .P., Elvevåg, B., Foltz, P. W., Tucker, R. P., **Holmlund, T. B.**, Cohen, A. S. (2018). Comparing Static and Dynamic Predictors of Risk for Hostility in Serious Mental Illness: Preliminary Findings. *Schizophrenia Research. 204*, 432-433. doi: 10.1016/j.schres.2018.08.030

Crow, T. (1998). Nuclear schizophrenic symptoms as a window on the relationship between thought and speech. *British Journal of Psychiatry, 173*(4), 303-309. doi: 10.1192/bjp.173.4.303

DeLisi, L. E., 2001. Speech disorder in schizophrenia: Review of the literature and exploration of its relation to the uniquely human capacity for language. *Schizophr. Bull. 27,* 481–496. doi: 10.1093/oxfordjournals.schbul.a006889

Deng, L., & Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(5), 1060-1089. doi: 10.1109/TASL.2013.2244083

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL]

Diener, C., Kuehner, C., Brusniak, W., Ubl, B., Wessa, M., & Flor, H. (2012). A meta-analysis of neurofunctional imaging studies of emotion and cognition in major depression. *NeuroImage, 61*(3), 677-685. doi: 10.1016/j.neuroimage.2012.04.005

Ebert, D., Van Daele, T., Nordgreen, T., Karekla, M., Compare, A., Zarbo, C., . . . Baumeister, H. (2018). Internet- and Mobile-Based Psychological Interventions: Applications, Efficacy, and Potential for Improving Mental Health. *European Psychologist, 23*(2), 167-187.

Elvevåg, B., Cohen, A. S., Wolters, M. K., Whalley, H. C., Gountouna, V. E., Kuznetsova, K. A., Watson, A. R., Nicodemus, K. K. (2016). An examination of the language construct in NIMH's Research Domain Criteria: Time for reconceptualisation! *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 6*(171), 909-919. doi: 10.1002/ajmg.b.32438

Elvevåg, B., Foltz, P. F., Rosenstein, M. & DeLisi, L. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics, 23*, 270-284. doi: 10.1016/j.jneuroling.2009.05.002

Elvevåg, B., Foltz, P. W., Rosenstein, M., Ferrer-I-Cancho, R., De Deyne, S., Mizraji, E., & Cohen, A. (2017). Thoughts About Disordered Thinking: Measuring and Quantifying the Laws of Order and Disorder. *Schizophrenia Bulletin, 43*(3), 509–513. doi:10.1093/schbul/sbx040

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

Elvevåg, B. & Goldberg, T. E. (2000) Cognitive impairment in schizophrenia is the core of the disorder. *Critical Reviews in Neurobiology, 14*, 1-21. doi: 10.1615/CritRevNeurobiol.v14.i1.10

Elvevåg, B., Weinstock, D. M., Akil, M., Kleinman J. E., & Goldberg, T. E. (2001). A comparison of verbal fluency tasks in schizophrenic patients and normal controls. *Schizophrenia Research, 51*(2), 119-126. doi: 10.1016/S0920-9964(00)00053-0

Faurholt-Jepsen, M., Brage, S., Vinberg, M., Christensen, E. M., Knorr, U., Jensen, H. M., & Kessing, L. V. (2012). Differences in psychomotor activity in patients suffering from unipolar and bipolar affective disorder in the remitted or mild/moderate depressive state. *Journal of Affective Disorders, 141*(2-3), 457-463. doi: 10.1016/j.jad.2012.02.020

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E. M., Bardram, J., & Kessing, L. V. (2014). Daily electronic monitoring of subjective and objective measures of illness activity in bipolar disorder using smartphones- the MONARCA II trial protocol: A randomized controlled single-blind parallel-group trial. *BMC Psychiatry, 14*(1), 309. doi: 10.1186/s12888-014-0309-5

Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E. M., Winther, O. . . Kessing, L. V. (2016). Voice analysis as an objective state marker in bipolar disorder. *Translational Psychiatry*, 6(7), E856. doi: 10.1038/tp.2016.123

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. (2002). Structured Clinical Interview for DSM–IV–TR Axis I Disorders–Patient edition. Retrieved from https://www.columbiapsychiatry.org/node/ 13821

Firth, J. R. (1957) *A synopsis of linguistic theory*. In *Studies in Linguistic Analysis*, pp. 1-32. Blackwell, Oxford

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: A meta-analysis of randomized controlled trials. *World Psychiatry, 16*(3), 287-298. doi: 10.1002/wps.20472

Friston, K., Frith, C., Liddle, P., & Frackowiak, R. (1991). Investigating a Network Model of Word Generation with Positron Emission Tomography. *Proceedings of the Royal Society B: Biological Sciences, 244*(1310), 101-106. doi: 10.1098/rspb.1991.0057

Fox, M., & Lobo, M. (2019). The molecular and cellular mechanisms of depression: A focus on reward circuitry. *Molecular Psychiatry*. doi: 10.1038/s41380-019-0415-3

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

Green, M. (1996). What are the functional consequences of neurocognitive deficits in schizophrenia? *American Journal of Psychiatry, 153*(3), 321-330. doi: 10.1176/ajp.153.3.321

Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia, 42*(9), 1212-1222. doi: 10.1016/j.neuropsychologia.2004.02.001

Hinton, G., Li, D., Dong, Y., Dahl, G. E., Mohamed, A., Jaitly, N., . . . Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 29(6), 82-97. doi: 10.1109/MSP.2012.2205597

Hoffman, P. (2019) Reductions in prefrontal activation predict off-topic utterances during speech production. *Nature Communications, 10*(1), 2041-1723. doi: 10.1038/s41467-019-08519-0

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S. & Elvevåg, B. (2019). Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry Research, 273,* 767-769. doi: 10.1016/j.psychres.2019.02.014

Holmlund, T. B., Fedechko, T. L., Elvevåg, B. & Cohen, A. S. (in press). Chapter 28: Tracking language in real time in psychosis. In: *A Clinical Introduction to Psychosis: Foundations for Clinical and Neuropsychologists*. Ed. J.C. Badcock & G. Paulik-White. Elsevier.

Holmlund, T. B., Foltz, P. W., Cohen, A. S., Johansen, H. D., Sigurdsen, R., Fugelli, P., ..., & Elvevåg, B. (2019a). Moving psychological assessment out of the controlled laboratory setting and into the hands of the individual: Practical challenges. *Psychological Assessment, 31*(3), 292-303. doi: 10.1037/pas0000647

Holmlund, T. B., Foltz, P. W., Cohen, A. S., Cheng, J., Bernstein, J., Rosenfeld, E., Elvevåg, B. (2019b) 24.4 Moving speech technology methods out of the laboratory: Practical challenges and clinical translation opportunities for psychiatry, *Schizophrenia Bulletin*, *45*(Issue Supplement_2), S129. doi: 10.1093/schbul/sbz022.099

Hsin, H., Fromer, M., Peterson, B., Walter, C., Fleck, M., Campbell, A., . . . Califf, R. (2018). Transforming Psychiatry into Data-Driven Medicine with Digital Measurement Tools. *npj Digital Medicine, 1*(1). doi: 10.1038/s41746-018-0046-0

Inkster, B., Stillwell, D., Kosinski, M., & Jones, P. (2016) A decade into Facebook: Where is psychiatry in the digital age? *Lancet Psychiatry, 3*(11), 1087-1090. Doi: 10.1016/S2215-0366(16)30041-4

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., ... Wang, P. (2010) Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry, 167*(7), 748-751. doi: 10.1176/appi.ajp.2010.09091379

Insel, T. R. (2017). Digital Phenotyping: Technology for a New Science of Behavior. *JAMA, 318*(13), 1215-1216. doi:10.1001/jama.2017.11295

Iter, D., Yoon, J., & Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, New Orleans, LA, USA, June (pp 136-146). doi: 10.18653/v1/W18-0615

Jablensky, A., Sartorius, N., Ernberg, G., Anker, M., Korten, A., Cooper, J., . . . Bertelsen, A. (1992). Schizophrenia: Manifestations, incidence and course in different cultures A World Health Organization Ten-Country Study. *Psychological Medicine. Monograph Supplement, 20*, 1-97. doi:10.1017/S0264180100000904

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *The American Journal of Psychiatry, 160*(1), 13-23. doi: 10.1176/appi.ajp.160.1.13

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., & Mcinnis, M. G. (2014). Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Florence, Italy, 4-9 May (pp. 4858–4862). doi:10.1109/ICASSP.2014.6854525

Kern, R., Green, M., Nuechterlein, K., & Deng, B-H. (2005). NIMH-MATRICS survey on assessment of neurocognition in schizophrenia. *Schizophrenia Research. 72*. 11-9. doi: 10.1016/j.schres.2004.09.004.

Kim, N., Kim, J.-H., Wolters, M. K., MacPherson, S. E., & Park, J. C. (2019). Automatic Scoring of Semantic Fluency. *Frontiers in Psychology, 10*(1020). Doi: 10.3389/fpsyg.2019.01020

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Maciver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron, 93*(3), 480-490. doi: 10.1016/j.neuron.2016.12.041

Krukow, P., Harciarek, M., Morylowska-Topolska, J., Karakuła-Juchnowicz, H., & Jonak, K. (2017). Ineffective initiation contributes to deficient verbal and non-verbal fluency in patients with schizophrenia. *Cognitive Neuropsychiatry, 22*(5), 391-406. doi: 10.1080/13546805.2017.1356710

Kuperberg, G. R., & Heckers, S. (2000). Schizophrenia and cognitive function. *Current Opinion in Neurobiology, 10*, 205-210. doi: 10.1016/s0959-4388(00)00068-4

Kuperberg G. R. (2010a). Language in schizophrenia Part 1: an Introduction. *Language and Linguistics Compass, 4*(8), 576–589. doi:10.1111/j.1749-818X.2010.00216.x

Kuperberg G. R. (2010b). Language in schizophrenia Part 2: What can psycholinguistics bring to the study of schizophrenia...and vice versa?. *Language and Linguistics Compass, 4*(8), 590–604. doi:10.1111/j.1749-818X.2010.00217.x

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, Cognition and Neuroscience*, *31*(1), 32–59. doi:10.1080/23273798.2015.1102299

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From Word Embeddings To Document Distances. *Proceedings of the 32nd International Conference on Machine Learning*, 37, Lille, France, 06-11 July (pp 957-966).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Le, T. P, Elvevåg, B, Foltz, P. W., **Holmlund, T. B.**, Schwartz, E. K., Cowan, T., & Cohen, A. S. (2018). Aggressive urges in schizotypy: Preliminary data from an ambulatory study. *Schizophrenia Research, 201*, 424-425. doi: 10.1016/j.schres.2018.05.045

Le T. P., Cowan T., Schwartz E. K., Elvevåg, B., **Holmlund T. B**., Foltz, P. W., ..., & Cohen, A. S. (2019). The importance of loneliness in psychotic-like symptoms: Data from three studies. Manuscript submitted for publication.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. doi: 10.1038/nature14539

Lezak, M. (2004). *Neuropsychological assessment* (4th ed.). Oxford: Oxford University Press.

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M., & Fuller-Tyszkiewicz, M. (2019). The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. *World Psychiatry, 18*, 325-336. doi: 10.1002/wps.20673

MacDonald, S. W., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences, 29*, 474-480. doi: 10.1016/j.tins.2006.06.011

MacDonald, S., Li, S-C., & Bäckman, L. (2009). Neural Underpinnings of Within-Person Variability in Cognitive Functioning. *Psychology and Aging. 24*. 792-808. doi: 10.1037/a0017798.

Marder S. R. (2006). The NIMH-MATRICS project for developing cognition-enhancing agents for schizophrenia. *Dialogues in Clinical Neuroscience, 8*(1), 109–113.

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

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 23, No. 23).

Marquand, A. F., Kia, S. M., Zabihi, M. K., Wolfers, T., Buitelaar, J., & Beckmann, C. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry, 24*, 1415-1424. doi: 10.1038/s41380-019-0441-1

Melle, I. (2019), Cognition in schizophrenia: a marker of underlying neurodevelopmental problems? *World Psychiatry, 18*: 164-165. doi:10.1002/wps.20646

Mikolov T., Chen K., Corrado G., & Dean J. (2013). Efficient estimation of word representations in vector space. In: *Workshop Proceedings for International Conference on Learning Representations 2013*. arXiv:1301.3781

Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia, 3*(1), 1-10. doi: 10.1038/s41537-017-0019-3

Mota, N. B., Sigman, M., Cecchi, G., Copelli, M., & Ribeiro, S. (2018). The maturation of speech structure in psychosis is resistant to formal education. *npj Schizophrenia, 4*(1), 25. doi: 10.1038/s41537-018-0067-3

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*(1), 182-191. doi: 10.1016/j.cortex.2013.12.004

Nielsen, J., & Landauer, T. K. (1993) A mathematical model of the finding of usability problems. In *CHI '93 Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, Amsterdam, Netherlands April 24 - 29 (pp 206-213). doi: 10.1145/169059.169166

Noel, V. A., Carpenter-Song, E., Acquilano, S. C., Torous, J., & Drake, R. E. (2019). The technology specialist: a 21st century support role in clinical care. *npj Digital Medicine*, *2*(1). doi: 10.1038/s41746-019-0137-6

Núñez, R., Allen, M., Gao, R., Miller, R. C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science? *Nature Human Behaviour, 3*(8), 782-791. doi: 10.1038/s41562-019-0626-2

Organization for the Review of Care and Health Applications (ORCHA) (2019). URL: https://www.orcha.co.uk

Pakhomov, S. V. S., Eberly, L., & Knopman, D. (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia, 89*, 42-56. doi: 0.1016/j.neuropsychologia.2016.05.031

Pal, R., Mendelson, J., Clavier, O., Baggott, M., Coyle, J., & Galloway, G. (2016). Development and Testing of a Smartphone-Based Cognitive/Neuropsychological Evaluation System for Substance Abusers. *Journal of Psychoactive Drugs, 48*(4), 288-294. doi: 10.1080/02791072.2016.1191093

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, A. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. (pp 1532-1543). URL: http://www.aclweb.org/anthology/D14-1162

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

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ..., & Vesely, K. (2011). The KALDI speech recognition toolkit, in *Proceedings IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii, USA, December, 2011.

Psychology Software Tools, Inc. (2016) E-Prime 3.0. Retrieved from https://www.pstnet.com.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019) Language Models are Unsupervised Multitask Learners. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In R. Witte, H. Cunningham, J. Patrick, E. Beisswanger, E. Buyko, U. Hahn, Hahn Verspoor, & A. Coden (Eds.), *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Framework.* Valletta, Malta, May (pp 45-50).

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia, 5*(1), 9. doi:10.1038/s41537-019-0077-9

Rosenstein, M., Foltz, P. W., DeLisi, L. E. & Elvevåg, B. (2015). Language as a biomarker in those at high-risk for psychosis. *Schizophrenia Research, 165*, 249-250. doi: 10.1016/j.schres.2015.04.023

Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology, 21*(4), 401-411. doi: 10.1037/0894-4105.21.4.401

Schlosser, D., Campellone, T., Truong, B., Etter, K., Vergani, S., Komaiko, K., & Vinogradov, S. (2018). Efficacy of PRIME, a Mobile App Intervention Designed to Improve Motivation in Young People With Schizophrenia. *Schizophrenia Bulletin, 44*(5), 1010-1020. doi: 10.1093/schbul/sby078

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shepherd, A., Sanders, C., & Shaw, J. (2017). Seeking to understand lived experiences of personal recovery in personality disorder in community and forensic settings – a qualitative methods investigation. *BMC Psychiatry, 17*(1), 282. doi:10.1186/s12888-017-1442-8

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J., & Botvinick, M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience, 40*, 99-124. doi: 10.1146/annurev-neuro-072116-03152

Silvert, W. (2001). Modelling as a discipline. *International Journal of General Systems, 30*(3), 261-282. doi:10.1080/03081070108960709

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of factorial experiments. R package version 0.19.1. https://CRAN.R-project.org/package=afex

Skelley, S. L., Goldberg, T. E., Egan, M. F., Weinberger, D. R., & Gold, J. M. (2008). Verbal and visual memory: characterizing the clinical and intermediate phenotype in schizophrenia. *Schizophrenia Research, 105*, 78-85. doi: 10.1016/j.schres.2008.05.027

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

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Bulletin, 30*(4), 415–433. doi: 10.1177/107769905303000401

Topol, E. (2019). *Deep medicine : How artificial intelligence can make healthcare human again* (First ed.). New York.

Torous, J., Anderson, G., Bertagnoli, A., Christensen, H., Cuijpers, P., Firth, J., . . . Arean, P.A. (2019). Towards a Consensus for Standards for Smartphone Apps and Digital Mental Health. *World Psychiatry, 18*(1), 97-98. doi: 10.1002/wps.20592

Vogt, H., Green, S., Ekstrøm, C. T., & Brodersen, J. (2019). How precision medicine and screening with big data could increase overdiagnosis. *BMJ, 366*, l5270. doi:10.1136/bmj.l5270

Waddell, G., Bircher, M., Finlayson, D., & Main, C. (1984). Symptoms and signs: Physical disease or illness behaviour? *British Medical Journal (Clinical Research Ed.), 289*(6447), 739-741. doi: 10.1136/bmj.289.6447.739

Ward, L. (2002). *Dynamical cognitive science*. Cambridge, Mass: MIT Press.

Watts, F., McKenna, F., Sharrock, R., & Trezise, L. (1986). Colour naming of phobia related words. *British Journal of Psychology, 77*(1), 97-108. doi: 10.1111/j.20448295.1986.tb01985.x

Wechsler, D. (1997). Wechsler Memory Scale - Third Edition, WMS-III: Administration and scoring manual. San Antonio, TX: The Psychological Corporation.

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

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLoS ONE, 11*(3), E0152719. doi: 10.1371/journal.pone.0152719

Wittgenstein, L., & Anscombe, G. (1967). *Philosophical investigations = Philosophische Untersuchungen* (3rd ed.). Oxford: Blackwell.

Wootton, R., Craig, J., & Patterson, V. (2006). *Introduction to telemedicine* (2nd ed.). London: Royal Society of Medicine Press.

Wolfers, T., Doan, N. T., Kaufmann, T., Alnæs, D., Moberget, T., Agartz, I., . . . Marquant, A. F. (2018). Mapping the Heterogeneous Phenotype of Schizophrenia and Bipolar Disorder Using Normative Models. *JAMA Psychiatry, 75*(11), 1146-1155. doi: 10.1001/jamapsychiatry.2018.2467

Wolfers, T., Beckmann, C., Hoogman, M., Buitelaar, J., Franke, B., & Marquand, A. (2019). Individual differences v. the average patient: Mapping the heterogeneity in ADHD using normative models. *Psychological Medicine*, 1-10. doi: 10.1017/S0033291719000084

World Health Organization (1995) Physical status: The use and interpretation of anthropometry. *WHO - Technical Report Series, 854*, vii-409. URL: https://www.who.int/childgrowth/publications/physical_status/en/

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

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]

Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2006). *The HTK Book Version 3.4*. Cambridge University Press.

Paper I.

Holmlund, T. B., Foltz, P. W., Cohen, A.S., Johansen, H., Sigurdsen, R., Fugelli, P., Bergsager, D., Cheng, J., Bernstein, J., Rosenfeld, E., & Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological Assessment, 31*(3), 292-303. doi: 10.1037/pas0000647

Paper II.

Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S., Bernstein, J., Rosenfeld, E., Laeng, B., & Elvevåg, B. (submitted). Using automated speech processing for repeated measurements of attentional bias and control. Manuscript under revision.

Using automated speech processing for repeated measurements of attentional bias and control

Terje B. Holmlund*, Department of Clinical Medicine, University of Tromsø, Norway;
Jian Cheng, Analytic Measures Inc, Palo Alto, California;
Peter W. Foltz, Institute of Cognitive Science, University of Colorado and Pearson PLC, London;
Alex S. Cohen, Department of Psychology, Louisiana State University;
Jared Bernstein, Analytic Measures Inc, Palo Alto, California;
Elizabeth Rosenfeld, Analytic Measures Inc, Palo Alto, California;
Bruno Laeng, Department of Psychology, University of Oslo, Norway;
Brita Elvevåg, Department of Clinical Medicine, University of Tromsø, Norway and Norwegian Centre for eHealth Research, University Hospital of North Norway, Tromsø, Norway

*Corresponding author: terje.holmlund@uit.no, Department of Clinical Medicine, UNN Åsgård, Postbox 6124, 9291 Tromsø, Norway

Abstract

Attentional bias and control are of critical importance to human behaviour. The gold standard measurement tool is the century old color-word Stroop interference task administered in laboratories in a cross-sectional manner. We investigated whether both traditional and novel metrics of interference could be obtained using automatic speech recognition on a spoken variant. We moved the task out of the laboratory to self-administration via smart devices, and ensured the design was suitable even for participants who by definition had attentional problems (143 participants; 86 healthy volunteers and 57 with psychiatric diagnoses). The interference effects were robust, and remained despite repeated testing. In addition to the traditional metrics of response onset latency, the duration of vocal utterances was derived and shown to be longer in those with a clinical diagnosis. This framework of remote assessment using speech processing technology enables the fine-grained longitudinal charting of attention.

Introduction

Words can grab our attention, and depending on the context and our experience, some words are more arresting than others. Noticing a sign with the word "Sharks" while swimming most definitely should capture our attention very rapidly and translate into a rapid escape response (Figure 1, panel a). However, in other contexts the automatic affective response associated with a word needs to be overridden such that, for example, when someone with a spider phobia reads the word "Spider-chart" in a work setting (Figure 1, panel b), the automatic urge to escape should be controlled. In these instances, the meaning of a word, and its corresponding internal representation, can have strong affective components resulting in behavioral change. Understanding these behavioral changes can lead to important clinical insights. The Stroop color-word interference task[1,2] is commonly regarded as the "gold standard" in research investigating attentional control and its biases.[3] Various stimulus word sets are employed, but since both context and personal concerns change rapidly, current standardized ways of testing likely miss important information. Individualized and customizable methods are needed, but this necessitates first establishing the proof-of-concept. We explore here how new technologies could be leveraged to maximize - and go beyond - the utility of the classic task.

The manner of speaking reveals our internal states (e.g., urgency or confusion), and a combination of recording technology and speech processing techniques enables the creation of detailed descriptions of **how** even single words are uttered. Healthcare professionals can do this intuitively and with great precision, detecting minute changes in emphasis or tone in utterances in the people they interact with (e.g., a hesitation in *"I feel ..... fine»,* or a longer, drawn out word in *"I am not reeeeally fine")*. These expressions, the prosodic elements of speech, are ephemeral, lost when the conversation is over, and documenting them is left to qualitative descriptions. Much can be gained by listening to the verbal behavior, and technology affords this in an objective and robust fashion. The Stroop task serves as a useful framework for exploring which speech signals are important. This may be useful in many circumstances, from aerospace to healthcare, and employing such techniques in psychopathology can serve as the ultimate test bed.

Several variants of the Stroop task leverage features of spoken responses. In the canonical example the task consists of naming the ink-color of a printed word, while ignoring the meaning of the word itself. If the printed word and its color are incongruent (e.g., the word **RED** printed in blue ink), the over-learned automated process of reading the word

interferes and produces a conflict cost, making response latencies (i.e., **when** something is said) slowed by 100-200 milliseconds as well as increasing the incidence of errors (i.e., the **what** is said), classically called *interference*. The task is easy to implement, can be fun to perform, and has gained a massive following and literature.[2] The Stroop task serves as a "model world" for measuring control over cognitive processes, as the rule (i.e., focus on color) provides a putative context to override a stimulus-driven, spontaneous response tendency of reading the printed word.[4] Put differently, the person doing the task needs to selectively attend to task-relevant dimensions and use the correct action strategy to correctly respond, also by recruiting the necessary cognitive resources for inhibitory control.[5] The task can also be useful for mapping attentional biases in for example depression, anxiety, post-traumatic stress disorder, phobias,[6] and schizophrenia.[7] Relevant to the aforementioned spider phobia, words such as "spider" and "crawl" can lead to 190 ms interference delays in affected individuals.[8] Being able to measure effects of this magnitude (i.e., 100-200 ms) in a practical manner is the first step towards real longitudinal tracking of attentional bias and control, but previous attempts using speech processing software have have reported a lack of sufficient temporal precision to yield quantifiable experimental effects.[9]

The dynamical nature - its stability and fluctuation - of cognition is of importance in the evaluation and monitoring of mental states, and thus has clinical, forensic, and educational significance. The early studies with a card version of the Stroop task relied on stopwatches for timekeeping of blocks of responses, thus any trial-to-trial effects were obscured. New technology makes it possible to obtain frequent and detailed measurements of the time-course of performance, as software can present stimulus words sequentially on a screen and automatically measure individual responses.[10,11,12,13] Such advances enable the fine-grained documentation of the individual differences in responses, as well as post-conflict activity and adjustments when the task is combined to physiological measures such as electroencephalography[14] or pupillometry,[15] creating even more direct assays of cognitive recruitment. Previous studies have also for the most part adopted a cross-sectional approach measuring trait-like differences between individuals. The digitalization behavioral tasks enables more frequent testing, and mobile devices additionally affords self-administration.[16] There have been other successful implementations of the Stroop test using mobile devices for data collection,[17,18,19] but these implementations involved responding by pressing on-screen buttons (e.g., buttons labeled "red", "blue" etc.). Using buttons complicates the response process by demanding a visual scan of the screen to locate the correct place to press, as well as the non-trivial processes involved in achieving a precise upper-limb motor action. In

addition to this there are technical challenges of making response time measurements related to different sampling rates of different device screens (usually 60Hz, i.e., ± 16 ms temporal resolution, may vary between devices). We will therefore argue that using speech as a medium provides the most direct assay of response performance and of the underlying cognitive processes.

We leveraged the Stroop test as a framework to address the following questions: (1) Can the classic attentional interference effects be measured in a self-administered speech-based test outside of the controlled laboratory settings, and is it - as expected - poorer in people who by definition have attentional problems? We expected it would be possible to measure response latencies (i.e., when responses were made) with a precision equal to lab-based methodologies, such that conflicting stimuli would result in slower and less accurate responses, and that performance would be slower, more variable and less accurate in a patient group. (2) Will this conflict cost decrease over time, or are there dynamic relationships in responses that can be leveraged to describe performance? We did not expect the interference effect to significantly weaken with practice over a period spanning five separate days, given how robust this effect has proven to be in numerous experimental settings. (3) Are there differences in how responses are uttered? We selected a measurement of the duration of the response utterance, from the onset of word to the following silence, and expected to find overall shorter utterances in the healthy group, courtesy of an expected higher incidence of slurred speech in patients. Whether or not this feature would be affected by word category interference was unknown. In sum, we sought to efficiently measure **when** something is spoken, **what** word is spoken, and **how** the word is uttered (Figure 1, panel d; results also reported in Table 1 as a function of **when**, **what** and **how**).

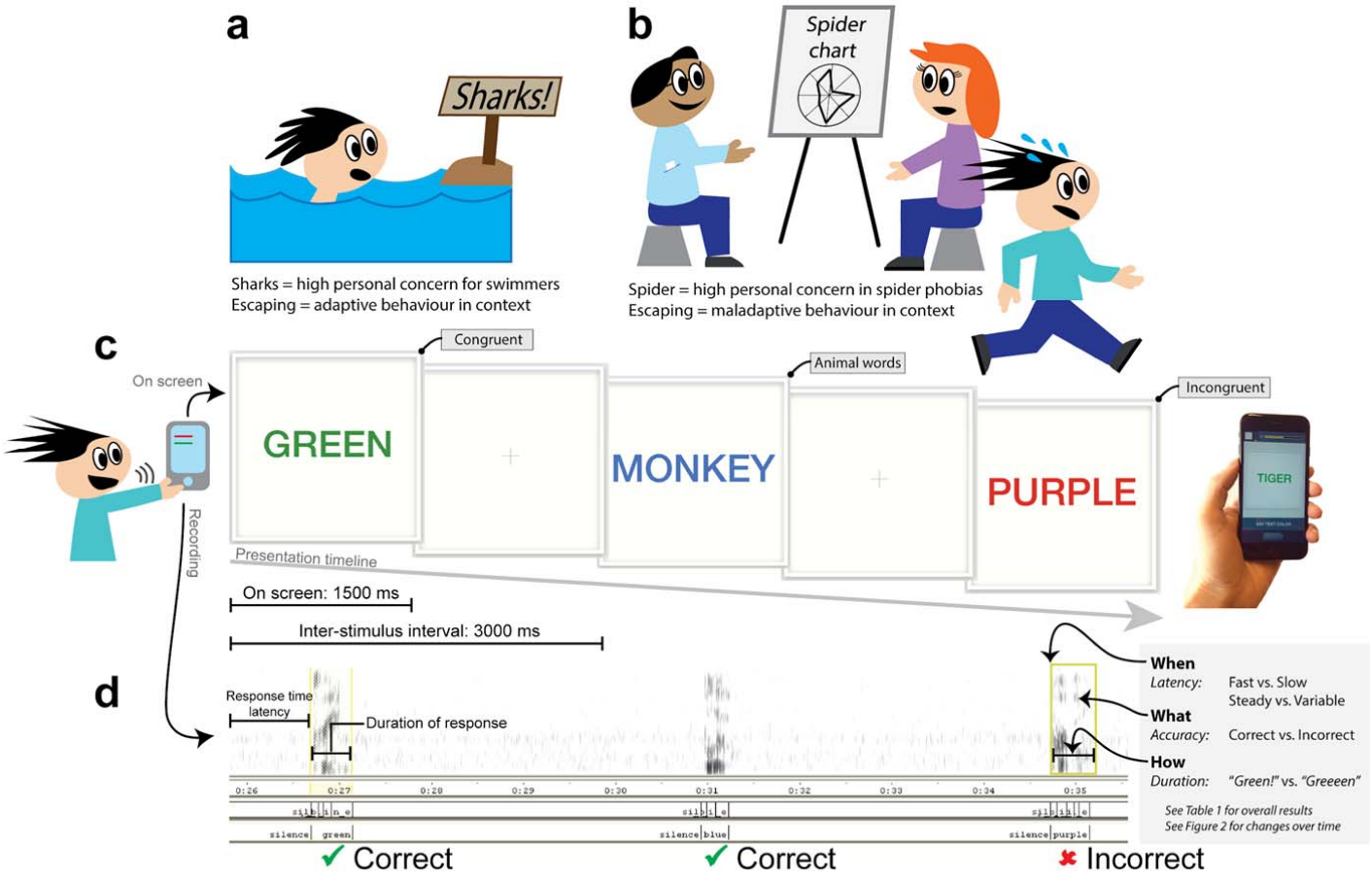


Figure 1: Words can invoke different biases of attention depending on context, and we present details on how behavioral data on word category effects can be collected using smart devices. *Panel a:* A situation where seeing the word "sharks" is of high personal concern and should promptly influence the swimmer to escape danger. *Panel b:* If an individual with an excessive fear of spiders encounters the word "spider", it may elicit an affective response. If the individual is not able to control attention and processing away from spider-related representations and decides to escape the situation (e.g., leave a meeting), it would constitute maladaptive behavior in a work environment. *Panel c:* Three different stimulus conditions were presented visually in a random order on the screen of a smart device, with a total of 32 presentations per testing session. In the example, first is presented a trial with the label "*congruent"*, where the word GREEN is printed in a green color. The word remained on the screen for 1500 milliseconds, followed by 1500 milliseconds with a fixation cross, before presentation of an "*animal word*" trial with the word MONKEY presented in blue color. Last is illustrated an "*incongruent"* trial, where the word PURPLE was presented in red, representing a conflict between ink color and the meaning of the word. *Panel d:* Spoken responses of naming ink colors were recorded, and automatic speech recognition software detected response latency, duration and accuracy. The file with recorded audio was segmented into either "silence" or the phonemes of the respective responses, making it possible to ignore phonations of hesitations such as "uh". The timestamp of when the stimulus word was flashed on the screen was subtracted from the word onset timestamp to measure the response time latency (the "**When**"). Responses were classified as either "Correct" or "Incorrect" (the "**What**"), and incorrect responses were not included in response time analysis. The "**How**" was indexed by the duration of spoken utterance (e.g., 'greeeen' versus 'green').

Results

To summarise the key findings, we found (i) clear differences between the different conditions (Table 1 & Figure 2, panel a), and thus demonstrated that this experimental operationalization of the Stroop interference task successfully replicated the classic effects, namely that the actual color word interfered with the naming of an incongruent ink color. While participants became faster over time these differences (ii) remained robust (the parallel trajectories in Figure 2 panel a and the stationary difference scores in Figure 2 panel d). The assays were (iii) suitably sensitive to group differences in speed and variability (Figure 2 panels b & c). Critically, the groups differed in the processing of animal words versus congruent color-words, which is in keeping with the existing literature but importantly extends this from the traditional version to a self-administered and spoken version that can be highly suited for adaptive testing purposes. Beyond this (iv) the duration of responses was clearly different between healthy participants and patients, which promises to be a useful metric in differentiating individual differences, a radically new approach courtesy of leveraging the state of the art automatic speech recognition with accurate timestamping of features in responses other than onset latency (Table 1; Figure 2, panel f). Although we only showcase in detail a few of the critical variables as proof of concept, we regard this as the roadmap towards a new research terrain where a multitude of acoustic features can be extracted from single word responses.

Table 1

	Healthy (N=86)		Patients (N=57)				
	M	SD	M	SD	d	t	p
When: Response Latency							
Speed (ms)							
Overall	751	66	810	96	0.71	4.01	0.001
Congruent	689	74	720	75	0.42	2.43	0.033
Animal words	736	63	805	104	0.79	4.43	<0.001
Incongruent	849	91	926	140	0.65	3.65	0.002
Stroop Effect scores (ms)							
Incongruent - Animals	113	53	121	62	0.14	0.83	0.407
Animals - Congruent	48	46	85	70	0.63	3.53	0.002
Variability (CV%)	20	3	22	4	0.69	3.93	0.001
What: Response Accuracy							
Accuracy (%)	94	7	86	12	-0.76	-4.21	<0.001
How: Response Duration							
Duration (ms)	493	62	556	63	1.01	5.9	<0.001

Note: Results express the between-persons means and standard deviations of scores derived from combining data from all five sessions. Differences between groups expressed in Cohen's d effect size ('d'), Welch's t-test t-value ('t') and statistical significance of t-test ('p') with Holm-correction for 12 tests. CV = Coefficient of Variation, in percentage.

Overall effects

To establish whether our paradigm was sensitive and robust enough to elicit the expected patterns of behaviour, we first examined all responses from participants in a combined fashion, the *personal scores*. This constituted 160 trials in total per participant, 40 trials in the *congruent* and *incongruent* conditions, 80 trials in the color-neutral *animal-words* condition. Table 1 therefore represents the group-wise means and standard deviations of scores based on the total set of responses from each participant. These were the most accurate scores that could be derived over a 5 day period with this paradigm. We expected word categories (i.e., *congruent* or *incongruent* color words or color-neutral *animal-words*) would affect performance, and that patients would generally perform poorer and specifically be more affected by word categories, indicating a reduction in attentional control.

When?- Response time latency. Processing speed, as represented by response time latency (henceforth RT), showed the expected pattern where responses were fastest in the congruent, non-conflict condition (Mean = 705 ms), slower in the animal words condition (Mean = 759 ms), and slowest in the incongruent, conflict condition (Mean = 870 ms; Table 1; Figure 1, panel a). A 2*3 repeated measures analysis of variance (henceforth rmANOVA) of response times with Group (healthy, patients) as the between-subject factor and Condition (*congruent, animal-words, incongruent*) as within-subject factors showed a significant main effect of Condition ($F(2,282) = 522.4$, $p < 0.001$).

The congruency-based interference measure, operationalized as the mean of incongruent responses minus *animal-word* responses, was on average 116 ms (SD = 57 ms). Consistent with previous single-trial studies in patients with mental illness (e.g., psychosis)[11,20] there was not a disproportionate *interference* effect when operationalized with *interference* difference scores ($d = 0.14$, $t = 0.83$, $p = 0.814$).

The difference between response latencies from stimuli with congruence between word meaning and ink color compared to the purportedly neutral *animal-words*, commonly termed the congruency-based *facilitation*, was modest with a mean of 62 ms (SD = 60 ms). Patients showed disproportionately larger effects and had a mean *facilitation* benefit of 85 ms (SD = 70 ms) versus healthy participants' 48 ms (SD = 46 ms, $d = 0.63$, $t = 3.53$, $p = 0.003$). Analysis of only the first session showed a large difference between groups ($d = 0.75$, $t = 4.0$, $p < 0.001$), and these findings are further explored below, where interesting differences in the time-course of this effect are discussed.

Processing efficiency (Table 1; Figure 2, panel f), as expressed by the RT variability (Coefficient of Variation of RTs in percentage, henceforth CV) did not differ between the

congruent (Mean CV = 18.9 %, SD = 3.8 %) and *incongruent* (Mean CV = 20.0 %, SD = 4.0 %) trials, but was marginally worse in *animal-words* trials (Mean CV = 17.2 %, SD = 3.7 %) resulting in a significant main effect of Condition in a 2*3 (Group*Condition) rmANOVA ($F(2, 282) = 29.4$, $p < 0.001$). Most importantly, healthy participants showed lower variability (Mean CV = 19.7 %, SD = 2.9 %), compared to patients (Mean CV = 22.1 %, SD = 3.9 %), confirmed by a significant main effect of Group ($F(1, 141) = 12.1$, $p < 0.001$).

In sum, this paradigm successfully elicited and measured the classic Stroop response patterns, as well as the expected performance profile of slower and more inaccurate responses in patients. Within the different word categories, the largest difference between groups was found in the *animal-words* condition (d = 0.79). This was also demonstrated by notable differences between groups when comparing performance in the facilitatory condition (congruence between word and color) to the naming of the color of animal words, traditionally intended to serve as a color-neutral and semantically neutral stimulus set. The findings of increased *facilitation* in patients are consistent with what has been previously reported.[11,21] It can be argued that the animal words in this context present a special kind of semantic interference. This would mean that the term "*facilitation*" here can be somewhat of a misnomer, and that the selection of non-color words can be leveraged for parsing of attentional control.

What? - Accuracy. The percentage of correct responses also showed the expected pattern where accuracy was higher in the non-conflict congruent condition, lower in the *animal-words* condition and lowest in the incongruent, conflicting condition. A 2*3 rmANOVA of response accuracy with Group (healthy, patients) as the between-subject factors and Condition (*congruent, animal-words, incongruent*) as within-subject factors showed a significant main effect of Condition ($F(2,282) = 40.5$, $p < 0.001$). As expected, healthy participants were generally more accurate than patients with a significant main effect of Group ($F(1,141) = 21.8$, $p < 0.001$). Patients' performance accuracy was disproportionately affected by condition (represented by a significant Group*Condition-interaction, $F(2,282) = 5.8$, $p = 0.003$). However, this disproportionality should be considered in light of ceiling effects in accuracy that were particularly evident in the healthy group. The size of the difference between the groups in overall accuracy was medium to large (d = 0.76).

How? - Measuring the duration of responses. In addition to enabling response time analysis via precise measurements of response onset latency, speech analysis provided a measure of the *duration* of response utterances (Figure 1, panel d). We found a large difference in word utterance duration between groups, with responses of shorter duration in

healthy participants (Mean = 491 ms, SD = 139 ms) as compared to the patients (Mean = 555 ms, SD = 135 ms, see Figure 2, panel g). This difference was unlikely to be related to general differences in RTs, as there was only a weak correlation between the two (Pearson's R = 0.028). Interestingly, there was a relationship between stimulus condition and response duration that mirrored the findings in our RT latency analysis. Durations of utterances showed a pattern where responses were of longer duration in the non-conflict *congruent* condition (Mean = 518 ms, SD = 141 ms), as compared to the *animal-words* condition (Mean = 516 ms, SD = 142ms) ms) and shortest in the *incongruent* conflict condition (Mean = 512 ms, SD = 137 ms). A 2*3 rmANOVA with Group (healthy, patients) as the between-subject factor and Condition (*congruent*, *animal-words*, *incongruent*) as within-subject factors revealed a significant main effect of Condition ($F(2,282) = 4.1$, $p = 0.018$). The main effect of group was highly significant ($F(1,141) = 35.5$, $p < 0.001$), and there was a marginally significant Group*Condition interaction ($F(2,282) = 3.1$, $p = 0.046$), courtesy of more prominent effects of conditions in patients. One speculation regarding these results is that conflicting stimuli may have pressurized participants to give shorter utterances to rid themselves of discomforts related to conflict resolution.

We did not explore these differences beyond this tantalizing observation as we acknowledge that they are modest as compared with the temporal resolution of the timestamping methods, i.e., ± 10 ms. An instrument error of this magnitude was considered sufficiently accurate as compared to other commonly employed instrumental setups for response time latency measurement (e.g., experiments using ordinary computer keyboards can be confounded by keyboard delay times that can vary between 11 and 73 ms[22]), but the small differences in utterance durations calls for care when concluding from speech recognition timestamps. However, the very fact that healthy and patient participant groups displayed such different results on this measure of vocal response duration was noteworthy, and merits a future study specifically designed to parse vocal response speed and vocal response duration, in combination with other acoustical parameters. In sum, the duration of utterance is a response feature that until now has been unexplored yet there are clearly tantalizing clues that it could provide critical performance assays and facilitate in the deconstruction of performance in attentional control and speech tasks.

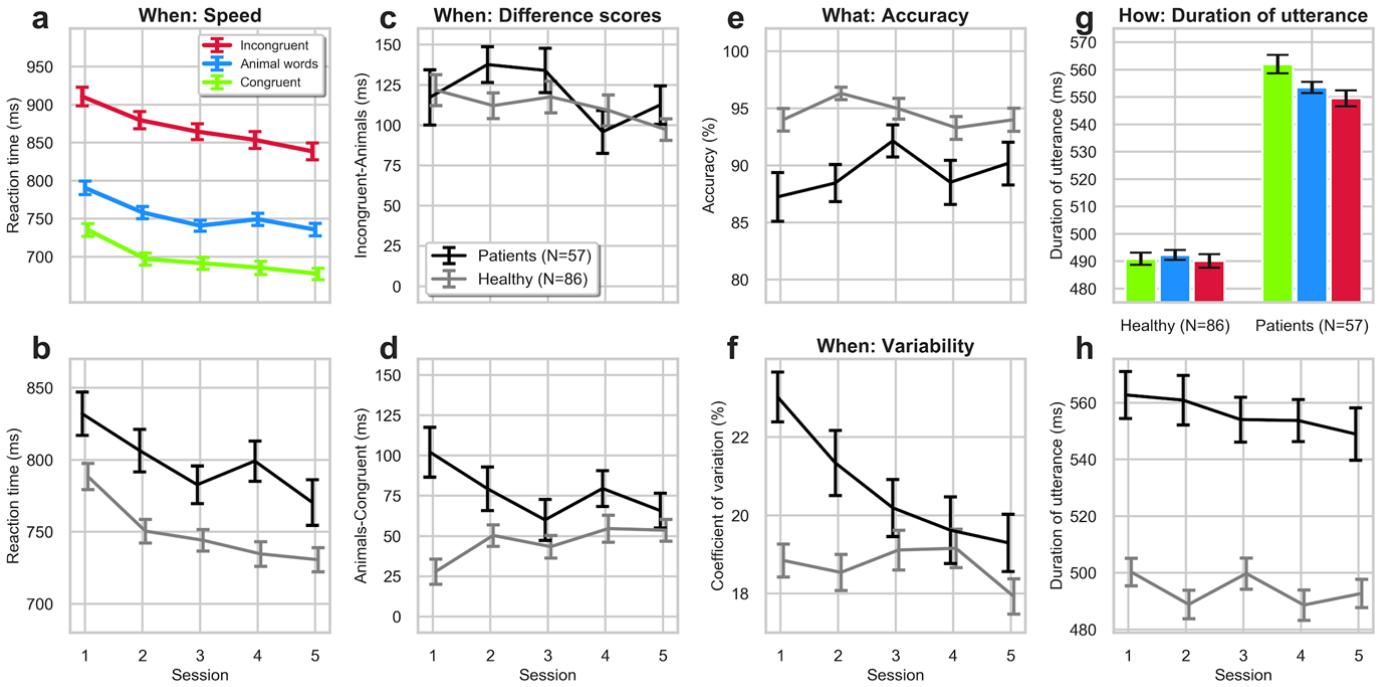


Figure 2. Tracking performance scores on the Stroop interference task over time. Error bars represent standard error of the mean. *Panel a:* There were large differences in response time speed between conditions, where longer response time latency in *incongruent* trials represents a cost of the conflict that needs to be resolved when the written word and its color do not match. As expected, participants got faster with practice, but the trajectories of the plots remained approximately parallel, representing the fact that the conflict effects did not extinguish over time. *Panel b:* Healthy participants were significantly faster to respond compared to patients, but groups improved comparably over time. *Panel c:* The difference between *incongruent* and *animal-word* trials was stable over time and similar between groups. *Panel d:* Groups were clearly different in the comparison between the *animal-word* responses and the non-conflict *congruent* color word responses. This may represent the level of interference by animal words, and was where there was a de-facto increase in attentional control in the patient group. *Panel e:* Healthy participant responded more accurately than patients, but patients improved accuracy with practice. *Panel f:* Response time variability, a sensitive measure of processing efficiency, demonstrated that the patient group improved gradually over time until they reached the level of stable performance of the healthy group. *Panel g:* The duration of vocal word utterances was shorter in healthy participants as compared to patients. Additionally, patient responses on *congruent* trials were of longer duration as compared to responses on the *incongruent*, conflict trials, but the differences were small and on the order of the resolution of the system (± 10 ms). These new metrics of vocal utterance duration may prove useful in parsing attentional bias and control in more customized paradigms. *Panel h:* The duration of response utterances was stable over time.

Performance over time

In order to investigate how performance changed over time, we calculated both measures of speed, intra-individual variability, accuracy and specific word category effects for each session, the *Session Scores*, for testing sessions occurring on five separate days. These were based on a few trials per calculation of performance metrics, in particular for the Session Stroop Effects, where for example the *Interference Effect* score would be based on the mean of 8 *incongruent* trials minus the mean of 16 *animal-words* trials. The overall variability of these scores were comparable to the *Personal Scores* that included a much larger sample of trials per score. *Personal Scores* for the overall response time speed had and overall SD of 84 ms, whereas the *Session Scores* had an SD of 92 ms. Measures of *Interference* (SD = 57 ms vs. SD = 87 ms) and *Facilitation* (SD = 60 ms vs. SD = 78 ms) also showed comparable variability.

In addition to these trends in between-person variability, the average within-person standard deviations of the Session Scores were of the same magnitude as the between-person standard deviations (Overall speed: Mean SD = 49 ms, *Interference*: Mean SD = 66 ms, *Facilitation*: Mean SD = 54 ms). Similar variability both within and between individuals provided an indication of ergodicity, making it more likely that insights gained from the group comparisons (e.g., that healthy participants perform better than patients) would be valid also at the individual participant level (e.g., that a participant would perform poorer when ill).[23] The fact that the *Session Scores* proved robust was extremely encouraging, since such minimalistic measurements represent data that can be reliably collected without sacrificing the usability of a remote assessment tool.

When? - Response latencies over time. As expected, with practice participants got faster (Figure 2, panel a). To formally examine this, we performed an rmANOVA with Group (healthy, patients) as the between-subject factors and Condition (*congruent, animal-words, incongruent*) and Session (sessions 1-5) as the within-subject factors. Speed increased over time (main effect of Session: $F(4, 524) = 26.7$, $p < 0.001$), similarly between groups (Group*Session interaction $F(4, 524) = 1.4$, $p = 0.25$, see Figure 2, panel b). There was a disproportionate increase in speed depending on conditions, resulting in a significant Group*Session*Condition interaction ($F(8, 1048) = 2.2$, $p = 0.023$). This disproportionality in how groups changed over time as a function of conditions is further explored under the analysis of *Interference* and *Facilitation*. Overall there was excellent test-retest reliability as indicated by the speed session scores (ICC = 0.91).

Interference scores were generally stable across sessions (Figure 2, panel c). For the *interference* measure (*Incongruent* RT - *Animal-words* RT) the main effect of Session was significant ($F(4, 524) = 2.95$, $p = 0.020$. This was likely due to the unusually low scores in patients in the fourth testing session. This was confirmed by a separate rmANOVA for patients with significant main effect of Session ($F(4, 192) = 2.93$, $p = 0.022$) and the largest difference between session three (Mean = 134 ms) and session four (Mean = 96 ms, $t = 2.75$, $p = 0.065$, Holm-corrected for 10 tests). The test-retest reliability of the interference scores for each session was lower compared to overall speed scores, but still moderate at ICC = 0.68.

For the *facilitation* measure (*Animal words* RT - *Congruent* RT; Figure 2, panel d) there was no overall main effect of Session ($F(4, 524) = 1.42$, $p = 0.228$), but the Group ($F(4, 1,131) = 9.71$, $p < 0.001$) and Group*Session interaction ($F(4, 524) = 5.68$, $p < 0.001$) were highly significant . Analyzing groups separately we see a main effect of Session in both groups (healthy: $F(4,332) = 3.31$, $p = 0.011$); patients: $F(4,192) = 3.36$, $p = 0.011$), but interestingly there was an opposite pattern where for healthy participants there was an increase in *facilitation* between session one (Mean = 28 ms) and session four (Mean = 55 ms, $t = 3.12$ $p = 0.019$; Holm-corrected for 10 tests), while in patients there was a decrease in facilitation between session one (Mean = 102 ms) and session three (Mean = 60 ms, $t = 3.36$, $p = 0.010$; Holm-corrected for 10 tests). Test-retest reliability was good (ICC = 0.79).

Practice also improved performance in patients as expressed by lower variability in RTs (Figure 2, panel f) over time, with the main effect of Session being highly significant ($F(4, 524) = 6.92$, $p < 0.001$), with a notable main effect of Group ($F(4, 524) = 10.8$, $p < 0.001$) and Group*Session interaction ($F(4, 524) = 5.18$, $p < 0.001$). Separate rmANOVAs for groups revealed a significant main effect of Session in patients ($F(4,192) = 7.38$, $p < 0.001$), but not in healthy participants ($F(4,332) = 1.73$, $p = 0.143$). These results indicated that it was those who by definition had attentional problems, namely patients who improved the most with practice, whereas healthy participants had already reached their optimal level of processing efficiency by the first session. Test-retest reliability for this score was moderate (ICC = 0.64).

Adopting a conservative approach when assessing the current findings we argue that the most robust difference between groups in this sample was that healthy participants (but not patients) displayed very little *facilitation* in the first testing session. This *facilitation* increased with practice, and can be interpreted as increasing semantic interference by the neutral condition *animal-words* over time, or alternatively that practice effects are disproportionately larger on the *congruent* trials creating a larger difference between *animal-*

words and *congruent* means. The reciprocal nature of the relationship between *interference* and *facilitation*, courtesy of the way these measures are operationalized in the current paradigm, presents a challenge when seeking to parse the processes that are putatively of different neurocognitive origins.[25]

What? - Accuracy over time. Overall, there was a tendency towards an improvement in accuracy over time across both groups (main effect of Session $F(4, 524) = 2.48$, $p = 0.043$), and the change over time was largest in the *incongruent* condition, with a Session*Condition interaction ($F(8,1048) = 2.26$, $p = 0.021$), most pronounced in patients as indicated by the significant Group*Session*Condition interaction ($F(8,1048) = 2.96$, $p = 0.003$). Accuracy retest reliability for sessions was good (ICC = 0.78).

How: Duration over time. Duration of responses did not change over time (main effect of Session $F(4, 524) = 0.60$, $p = 0.664$), indicating that how words were pronounced were more trait-like in their presentation, not so much state-dependent and affected by practice with the task. Test-retest reliability was also good (ICC = 0.76).

Discussion

We found that when administering the classic Stroop interference task remotely via smart devices it was possible to measure robust effects of word categories on color naming, with the expected pattern of delayed responses in the color-conflict condition, using speech recognition software. Healthy participants had faster, less variable and more accurate responses as compared to patients. There was no difference between groups on the traditional Stroop *Interference* measure, but patients did show larger *Facilitation* effects, a difference most pronounced in the first session. Speech processing tools additionally revealed differences in word utterance duration, where response word utterances were generally longer in patients. Furthermore, examination across multiple days revealed that even though the word category effects were robust - in the sense that they were not extinguished by practice - the assays were differentially moderated by practice and could potentially be used to uncover individual differences and differentiate clinical groups. Excitingly, these sizes of the measured effects are similar in magnitude to previous findings of clinical relevance (i.e., 50-400 ms interference by words of affective salience[6]), providing a proof-of-concept for future mobile remote administration that can leverage voice to assay and deconstruct attentional control and bias. The well-established Stroop paradigm therefore appears to be well suited as

a flexible and scalable platform for future investigations using smart devices and fast internet-based analysis and feedback.

Two different mechanisms in the classic Stroop model presented by Cohen et al.[4] can provide explanations for different trends in Stroop scores between groups: Patients' speed and accuracy increased over sessions, demonstrating a *de facto* increase in attentional control with reduction in semantic interference of animal words. The healthy volunteers, on the other hand, performed with very low levels of semantic interference from the beginning. With repeated exposure to the stimuli, a strengthening of the input units for animal words to the proposed attentional control network, possibly led to increased salience of these words and increased semantic interference. Such an interpretation is strengthened by the fact that the largest difference in response time latency between groups was found in the purportedly neutral condition (Table 1, row for "Animal words"). The ability to ignore irrelevant features, as well as the resting levels of inputs to the feature processing and behavioral output, converged between groups over time.

Several issues with the design of the current experiment may limit the generalization of findings beyond the demonstration of the effectiveness of speech processing technologies. First, in an idealized experiment one would prefer a significantly larger number of trials per participant. However, during the development of the assessment tool it quickly became evident that there was an optimal length of a session and individual tasks that ensured people would even use the remote self-administered system. Put differently, adding more trials would not be feasible from a usability perspective, and ultimately lead to less data. Second, the short time-span (i.e., five days) was insufficient to provide intra-individual comparison between periods with disordered states (e.g., psychosis, mania) versus stable states. With more time-points the longitudinal nature of our data could be more suited for robust examination using linear mixed effects modelling, latent change scores and latent growth curve models.[25] Even so, combining data from a participant over five days should be highly suitable to measure differences on a week-to-week basis. For example, it would allow for robust measurements of potential differences in performance before and after initiating a pharmacological intervention, or comparison between clinically stable phases for a patient in an outpatient setting versus when the same patient is hospitalized due to relapse. Third, this type of stimulus set, where both attentional control and attentional bias due to the salience of words affects performance, presents a complicated situation with many degrees of freedom for interpretation, but it also reveals the potential for suitable paradigms to parse both cognitive abilities and personal levels of word salience. The design in the current proof-of-

concept study employed a generic and well-explored set of stimuli, but the technology allows for vastly more complex, tailored and adaptive approaches.

Several of the limitations can be addressed with further development in task design where tailored procedures may provide more sensitive assays. Given that the most sensitive measures of word category effects were related to the putative neutral *animal-words* condition, there may be other categories of words that can more effectively parse the level of attentional control and bias in an individual. Arguably, animal words will have different affective components to different individuals, depending on their experiences, be that with dogs, monkeys or tigers. Moreover, putative color distances between the presented hues and those implied by the words' meaning could be controlled and carefully balanced, since previous studies have shown a role of such input factors in modulating the Stroop effect.[26] Ultimately, the current 'one-paradigm-fits-all' approach may not be sufficiently effective, and future methods could employ personalized adaptive paradigms able to tailor stimulus materials to more effectively gauge individual levels of performance and longitudinal change. Such adaptive paradigms may also be configured to be more entertaining to the user, thus allowing more trials and more robust metrics.

The demonstration of differences in word utterance durations holds promise of a multitude of new ways one can extract information from responses in spoken assessment tasks. Response properties are not limited anymore to simple accuracy and time-stamping measurements, as it is now technologically feasible to assay expression of affective states using prosodic elements of speech such as sound pressure and pitch. Indeed, we have previously found that acoustic variables derived from this seemingly innocuous Stroop task were remarkably more direct assays of affective states as compared to when such measures were derived from story retelling, picture description and even verbal self-reports on subjective state (i.e., "How do you feel today?"[27]). Put differently, affect measures derived from a person's utterance of a color word can provide crucial and clinically relevant signals in an inherently non-threatening manner, in that confrontation of potentially arousing or debilitating topics can be avoided. Naturally, acoustic metrics of affective states can provide a more complete picture of the neurocognitive state of the individual, as emotional valence and levels of arousal can have a modulating effect on cognitive performance.[28] Additionally, this can be expanded by combining the method with other objective measures, as the Stroop test is ideally set up for using pupillometry as a biomarker for arousal[29] and task demands or mental effort.[15] By mapping the individual distribution of performance and relevant biomarkers over

time, these technologies can enable us to assess the dynamic effects of emotional states in cognitive functions.

In conclusion, we have shown that an adaptation of a brief spoken Stroop paradigm implemented on a smart device can provide an experimental framework to enable the identification of specific attentional biases and assessment of the ability to control behaviour. These functions at the core of most cognitive processes critical for our everyday life. The methodology utilized in this study, both in terms of stimulus presentation and vocal response processing opens up new venues of longitudinal behavioral assessments in humans. Indeed, technology is changing the nature of behavioral assessment and research,[30] and the resulting models of brain function and dysfunction, and brings the promise of personalized medicine closer to realization.

Methods

A behavioral assessment task was developed that was similar in form and structure to the standardly employed single-trial Stroop, but specifically designed for daily and remote administration. We collected data from voice responses in a series of self-administered interactions over periods of five days with a smartphone-based testing application. This resulted in a total of 32692 responses collected in 1065 testing sessions. Sessions with the device lasted around 15 minutes and contained different tasks as part of a larger study on the assessment of language, memory and psychomotor skills, as well as self-report on mental states (see [16] for an overview of the tasks).

Participants

From a total of 224 participants who were tested with this Stroop task, we analysed responses from the subset of participants (N = 141, 63%) who completed five sessions. In this sample we compared the performance of 84 university students (19.8% male, mean age = 20.0, SD = 1.9) to the performance of 57 male inpatients (Mean age = 39.1, SD = 11.2) undergoing treatment for substance abuse disorders. Students were presumed to be relatively healthy and henceforth we refer to them as “healthy participants”. Patients had a primary diagnosis of substance abuse, most prevalently with addiction to alcohol (26%), cocaine (26%) and opioids (25%), additionally, 63% had psychiatric comorbidity, most commonly depression (39%). In light of the notable differences in health and age between healthy and

patient participants, we assumed that there would be differences in performance between groups, and therefore a valid measure of attentional control should reveal such a difference (i.e., part of the proof of concept).

The study was approved by the Louisiana State University Institutional Review Board (#3618), and all participants signed consent forms prior to participation. Students were rewarded with course credits for participation, while patients were given monetary rewards of $5 per completed session.

Procedure

Participants were asked to give verbal and touch-screen responses presented on a smart device using an in-house developed mobile application for the iOS operating system from Apple Inc. Each session with the smart device contained one sequence of Stroop task trials. A visual prompt appeared before the sequence commenced, with the words: "SAY TEXT COLOR", and a vocal prompt saying "Say the color the word is printed in". The first word presentation was initiated by the press of a touch-screen button from the user, then all subsequent presentations for the session appeared consecutively in a randomized sequence for 96 seconds. The paradigm was based on a well-established procedure introduced by Carter, Robertson and Nordahl [10] and used in numerous other studies (e.g., [11,12,13]). For the mobile implementation we made some notable adjustments. First, the number of trials was reduced to strike a balance between what would be an acceptable duration for an ambulatory task for chronically ill patients and what could produce a sufficiently high number of responses for statistical analysis. The usability aspects of the test development were of critical importance to achieving compliance from participants, as we received feedback during preceding experiments in the study from participants indicating that the duration of testing may have been too long. Second, we increased the pace of the task due to feedback preceding the study proper from users that the task was “sluggish”.

Thirty-two words were presented in three stimulus conditions (8 *congruent* stimuli, 8 *incongruent* stimuli and 16 animal-word stimuli). Congruent stimuli consisted of color-words printed in the same color that they represent e.g. “RED” printed in the red color. Incongruent stimuli consisted of color-words printed in one of the remaining three colors (e.g. RED printed in green color). For measurement of performance unrelated to color-word congruence, animal words of 3-6 letters (DOG, BEAR, TIGER, MONKEY) were presented in all of the four colors. Words were presented on a white background in capital letters (Arial bold font, height = 165 pixels) using four different colors: **RED**, **BLUE**, **GREEN** and **PURPLE**.

Words remained on the screen for 1500 ms, followed by a fixation cross for 1500 ms, resulting in a regular Inter-Stimulus Interval (ISI) of 300 ms (Figure 1, panel a). All responses recorded within the ISI were defined as a response to the preceding word, and responses after the ISI were thus defined as responses to the next trial.

Analysis

Speech recognition. Audio responses were recorded continuously throughout the Stroop task by the microphone built into the smart device, sampled at 16000 Hz and saved in a .flac-format for further processing (Figure 1, panel d). Voice response onsets were automatically timestamped at 10 milliseconds (ms) increments by an in-house developed automatic speech recognition model, using the Kaldi speech recognition toolkit.[31] Stimulus on-screen onset was also time-stamped, and the response latency was derived by calculating the duration between stimulus- and response timestamps. The language model was specifically tuned to recognize the relevant words in the Stroop task (i.e., the color words). This technique allowed us to take advantage of knowledge about the context of the spoken utterances, namely that words such as “GREEN” and “RED” were more likely to occur than “Car” or “Spoon”, thus increasing the accuracy of the word recognition. Performance was evaluated by comparing machine transcripts to 175 manually transcribed recordings and word error rate for the recognizer was calculated at 6.26%. This was considered highly accurate and determined to be acceptable.

As a consequence of using automatic speech recognition for detecting responses, conclusions regarding the presence and accuracy of responses may be confounded by processes of the recognizer. If there was no response detected, this may have been due to no utterance being made, but it may also have been that the utterance was indistinguishable from background noise (i.e., utterance was too weak or unclear to be detected as a word). Equally, a response detected as “incorrect” by the automated system may in fact be due to an incorrect word uttered (e.g. “RED” or “TIGER”, when correct is “GREEN”), but it may also be due to an automatic speech recognition error (e.g. the correct utterance “GREEN” is recognized as “BLUE”) due to the way it is pronounced, registering falsely as an error. To have an accurate response detected the response must be (i) the correct color word and (ii) clearly stated. Accuracy was then defined as: (Number of correct responses detected) / (Total number of presentations). It is acknowledged that this approach is extremely conservative such that responses from participants with slurred or otherwise impeded speech could be excluded from this particular analysis.

Data cleaning. Only responses recognized as "Correct" by the automatic system were included for response time (RT) processing. To limit the effect of artifactual outliers, we removed responses of less than 200 ms, as these were considered task-unrelated behaviour. Furthermore, we removed responses with latencies more than three standard deviations over the general mean for each group, separately for the three conditions, to avoid removing disproportionately more responses from the patient populations and from the conflict trials, known to have slower response times. Using three standard deviations as a cut-off is a widely employed method to avoid introducing biases in means due to the skewed distribution of response time data.[32] Overall, 1.8 % of trials were removed for patients and 1.6 % of trials for healthy participants. Lastly, sessions with less than four registered responses for any condition (i.e. < 50% accuracy in the condition) were not subject to further processing of session response time statistics (means, SDs and the Stroop effects).

Performance analysis. In order to extract a detailed description of response patterns on the Stroop task, we derived general metrics of performance alongside the conflict-related metrics that specifically assay attentional control. General performance metrics were processing *speed*, as measured by RT latencies (in milliseconds after stimulus onset), processing *efficiency* and central nervous system integrity as measured by intra-individual variability of RTs,[33] operationalized here as the SD of RT divided by mean of RT and expressed as a percentage, i.e., the coefficient of variation of RTs,[34] and *accuracy* (as percentage correct responses).

The specific goal of using the Stroop paradigm was to assess how conflicts between presented colors and word categories (i.e., *congruent*/*animal-words*/*incongruent*) affect the latency of responses. In order to deconstruct the different aspects of word processing, we calculated word category conflict effects in two related measures: (i) *Interference*, expressed as the difference between the mean response time of the *incongruent* trials and the mean response time on purportedly neutral trials, and (ii) *Facilitation*, expressed as the difference between the mean response time of the *congruent* trials and the mean response time of the color-neutral *animal-words* trials. Word category effect scores were also calculated as a ratio of overall speed, and the differences between conditions were divided by the mean response time across conditions, providing an individually calibrated measure of word category effects. Similar metrics have previously been employed (e.g., [11]), and the individualized calibration ensures a fairer metric with which to compare groups that have inherently different baselines in terms of overall processing speed. The differences between groups in ratio scores were

somewhat smaller between groups but did not affect conclusions and are thus not reported here.

Traditionally the focus of Stroop response analysis has been on *when* responses occurred (i.e., latencies), but digital recordings also allow for analysis of *how* a spoken response is uttered (i.e., acoustic properties). A variety of features are possible to extract from recordings (see [27]), but we demonstrate this concept by measuring response word duration, namely how long it takes from the start of an utterance to the subsequent silence (Figure 1, panel c).

We calculated two different levels of performance scores; *Personal Scores* and *Session Scores*. The *Personal Scores* combined all the trials a participant had done, across the different sessions, and should therefore provide a basis for the most robust assessment of performance. The *Session Scores* were derived from trials within one session, providing the best estimate of performance that could be derived on a single day. This distinction is important because the different levels of scores would promise different temporal resolutions of performance assessment: *Personal Scores* (e.g., combined over five days) could provide week-to-week assessment of function, while *Session Scores*, if robust, could provide useful information for day-to-day dynamics of mental states.

Statistical methods. The statistical significance between groups and conditions (i.e. present or not) was assessed with repeated measures analysis of variance (rmANOVA) performed with the Analysis of Factorial EXperiments package, implemented in the R programming language.[35] A broad exploration using post-hoc t-tests was conducted, and as such, marginally significant differences should only be considered suggestive. Results and degrees of freedom are expressed without corrections where Mauchly's test of sphericity showed unequal variances, but corrected results were examined and did not affect conclusions. The distribution of the resulting RT data was as expected non-normal and ex-Gaussian, but we nonetheless considered parametric tests appropriate (and analyses of log-transformed, standardized response times were additionally performed but did not affect conclusions). Test-retest reliability across the five sessions was assessed with intraclass correlations (ICC, absolute agreement) using the R-package "psych".[36]

Data availability

A de-identified subsample of the data will be available at https://opendata.uit.no, along with code illustrating the core analysis procedure.

Acknowledgements

This research was funded by the Research Council of Norway to Brita Elvevåg (#231395). We would like to thank Taylor L. Fedechko, Tovah M. Cowan and Thanh P. Le for their invaluable efforts in data collection and analysis of other parts of the project.

Competing Interests

The authors declare that there are no competing interests.

Author Contribution

All authors contributed to the design, analysis, interpretation and writing of the manuscript. The final version of the manuscript has been approved by all authors.

References

1. 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

2. 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

3. MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121*(1), 12-14. doi:10.1037/0096-3445.121.1.12

4. Cohen, J.D., Dunbar, K., & McClelland, J.L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332–361. doi: 1990-27437-001

5. Kahneman, D. (1973). Attention and effort (Prentice-Hall series in experimental psychology). Englewood Cliffs, N.J: Prentice-Hall.

6. Williams, J. M., Mathews, A., & MacLeod, C. (1996). The Emotional Stroop Task and psychopathology. *Psychological Bulletin, 122*(1), 3-24. doi: 10.1037/0033-2909.120.1.3

7. Henik, A., & Salo, R. (2004). Schizophrenia and the Stroop Effect. *Behavioral and Cognitive Neuroscience Reviews, 3*(1), 42-59. doi: 10.1177/1534582304263252

8. Watts, F., McKenna, F., Sharrock, R., & Trezise, L. (1986). Colour naming of phobia□related words. *British Journal of Psychology*, *77*(1), 97-108. doi: 10.1111/j.2044-8295.1986.tb01985.x

9. Pal, R., Mendelson, J., Clavier, O., Baggott, M., Coyle, J., & Galloway, G. (2016). Development and Testing of a Smartphone-Based Cognitive/Neuropsychological Evaluation System for Substance Abusers. Journal of Psychoactive Drugs, 48(4), 288-294. doi: 10.1080/02791072.2016.1191093

10. Carter, C. S., Robertson, L. C., & Nordahl, T. E. (1992). Abnormal processing of irrelevant information in chronic schizophrenia: Selective enhancement of Stroop facilitation. *Psychiatry Research, 41*(2), 137-146. doi: 10.1016/0165-1781(92)90105-C

11. 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

12. Barch, D. M., Carter, C. S., Perlstein, W. D., Baird, J., Cohen, J., & Schooler, N. (1999). Increased Stroop facilitation effects in schizophrenia are not due to increased automatic spreading activation. *Schizophrenia Research, 39*(1), 51-64. doi: 10.1016/S0920-9964(99)00025-0

13. Barch, D. M., & Carter, C. S. (2005). Amphetamine improves cognitive function in medicated individuals with schizophrenia and in healthy volunteers. *Schizophrenia Research, 77*(1), 43-58. doi: 10.1016/j.schres.2004.12.019

14. Schack, B., Chen, A. C. N, Mescha, S., & Witte, H. (1999). Instantaneous EEG coherence analysis during the Stroop task. *Clinical Neurophysiology, 110*(8), 1410-1426. doi: 10.1016/S1388-2457(99)00111-X

15. Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing, 12*(1), 13-21. doi: 10.1007/s10339-010-0370-z

16. Holmlund, T. B., Foltz, P. W., Cohen, A. S., Johansen, H. D., Sigurdsen, R., Fugelli, P., . . . Elvevåg, B. (2019). Moving psychological assessment out of the controlled laboratory setting: Practical challenges. *Psychological Assessment, 31*(3), 292-303. doi: 10.1037/pas0000647

17. Bajaj, J.S., Heuman, D.M., Sterling, R.K., Sanyal, A.J., Siddiqui, M.S., Matherly, S.C., . . . & Wade, J.B. (2015). Validation of EncephalApp, Smartphone-Based Stroop Test, for the Diagnosis of Covert Hepatic Encephalopathy. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association, 13,* 10, 1828-1835.e1. doi: 10.1016/j.cgh.2014.05.011

18. Spanakis, P., Jones, A., Field, M., & Christiansen, P. (2019). A Stroop in the Hand is Worth Two on the Laptop: Superior Reliability of a Smartphone Based Alcohol Stroop in the Real World. *Substance Use & Misuse, 54*(4), pp.692–698. doi: 10.1080/10826084.2018.1536716

19. Waters, A. J., & Li, Y. (2008). Evaluating the utility of administering a reaction time task in an ecological momentary assessment study. *Psychopharmacology, 197*, 25–35. doi: 10.1007/s00213-0071006-6

20. 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

21. Phillips, M.L., Woodruff, P.W.R, & David, A.S, 1996. Stroop interference and facilitation in the cerebral hemispheres in schizophrenia. *Schizophrenia Research, 20*(1), 57–68. doi: 10.1016/0920-9964(95)00088-7

22. Shimizu, H. (2002). Measuring keyboard response delays by comparing keyboard and joystick inputs. *Behavior Research Methods, Instruments, & Computers, 34*(2), 250-256. doi: 10.3758/BF03195452

23. Fisher, A., Medaglia, J., & Jeronimus, B. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America, 115*(27), E6106-E6115. doi: 10.1073/pnas.1711978115

24. Brown, T. L. (2011). The relationship between Stroop interference and facilitation effects: statistical artifacts, baselines, and a reassessment. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 85-99. doi: 10.1037/a0019252

25. Woodard, J.L. (2017). A quarter century of advances in the statistical analysis of longitudinal neuropsychological data. *Neuropsychology, 31*(8), 1020-1035. doi: 10.1037/neu0000386

26. Laeng, B, Låg, T., & Brennen, T. (2005). Reduced stroop interference for opponent colors may be due to input factors: Evidence from individual differences and a neural network simulation. *Journal of Experimental Psychology: Human Perception and Performance, 31*(3), 438-452. doi: 10.1037/0096-1523.31.3.438

27. Cheng, J., Bernstein, J., Rosenfeld, E., Foltz, P. W., Cohen, A. S., Holmlund, T. B. & Elvevåg, B. (2018). Modeling self-reported and observed affect from speech. In

Proceedings Interspeech, Hyderabad, India, 2-6 September (pp 3653-3657). doi: 10.21437/Interspeech.2018-2222

28. Weinbach, N., Kalanthroff, E., Avnit, A., & Henik, A. (2015). Can arousal modulate response inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1873-1877. doi: 10.1037/xlm0000118

29. McGarrigle, R., Dawes, P., Stewart, A., Kuchinsky, S., & Munro, K. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology, 54*(2), 193-203. doi: 10.1111/psyp.12772

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

31. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The KALDI speech recognition toolkit, in *Proceedings IEEE 2011 Workshop on Automatic Speech Recognition and Understanding,* Hawaii, USA, December, 2011.

32. Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 43*(4), 907-912. doi: 10.1080/14640749108400962

33. MacDonald, S.W.S, Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences, 29*(8), 474-480. doi: 10.1016/j.tins.2006.06.011

34. West, R., Murphy, K. J., Armilio, M. L., Craik, F. I., Stuss, D.T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain and Cognition, 49*(3), 402-419. doi: 10.1006/brcg.2001.1507

35. Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of factorial experiments. R package version 0.19.1. https://CRAN.R-project.org/package=afex

36. Revelle, W. (2018). psych: Procedures for psychological, psychometric, and personality research. R package version 1.8.10. https://CRAN.R-project.org/package=psych

Paper III.

Holmlund, T. B., Cheng, J.,Foltz, P. W., Cohen, A. S., & Elvevåg, B. (2019). Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry Research. 273*, 767-769. doi: 10.1016/j.psychres.2019.02.014

Paper IV.

Holmlund, T. B., Chandler, C., Foltz, P. W., Cohen, A. S., D., Cheng, J., Bernstein, J., Rosenfeld, E., & Elvevåg, B. (submitted). Applying speech technologies to assess verbal memory in patients with serious mental illness. Manuscript submitted for publication.

Applying speech technologies to assess verbal memory in patients with serious mental illness

Terje B. Holmlund
University of Tromsø, Norway
Chelsea Chandler
University of Colorado Boulder
Peter W. Foltz
University of Colorado Boulder
Pearson PLC, London, England
Alex S. Cohen
Louisiana State University
Jian Cheng, Jared C. Bernstein and Elizabeth P. Rosenfeld
Analytic Measures Inc, Palo Alto, California
Brita Elvevåg
University of Tromsø, Norway
Norwegian Centre for eHealth Research, Norway

Abstract

Verbal memory deficits are one of the most profound neurocognitive deficits associated with schizophrenia and serious mental illness in general. As yet, their measurement in clinical settings is limited to traditional tests that allow for limited administrations and require substantial resources to deploy and score. Therefore, we developed a digital ambulatory verbal memory test with automated scoring, and repeated self-administration via smart devices. 104 adults participated, comprising 25 patients with serious mental illness and 79 healthy volunteers. The study design was successful with high quality speech recordings produced to 92% of prompts (Patients: 86%, Healthy: 96%). The story recalls were both transcribed and scored by humans, and scores generated using natural language processing on transcriptions were comparable to human ratings ($R = 0.83$, within the range of human-to-human correlations of $R = 0.73$-0.89). A fully automated approach that scored transcripts generated by automatic speech recognition produced comparable and accurate scores ($R = 0.82$), with very high correlation to scores derived from human transcripts ($R = 0.99$). This study demonstrates the viability of leveraging speech technologies to facilitate the frequent assessment of verbal memory for clinical monitoring purposes in psychiatry.

KEY WORDS:
Memory assessment, natural language processing, automatic speech recognition, word embeddings, remote assessment, mobile devices

Introduction

Our ability to remember stories we have heard can be affected by conditions that affect cortical function. Specifically, the verbal processing component of episodic memory is a useful endophenotype in schizophrenia, with patients displaying a disproportionate impairment in verbal relative to visual episodic memory.[1,2,3] Indeed, verbal memory assessment is core to virtually every neuropsychological test battery for schizophrenia and for evaluating pharmacological and remediation-based interventions. Unfortunately, verbal episodic memory is traditionally assessed by counting units of information recalled, which requires trained personnel, and limits the tests' operationalization of *what memory actually is* (i.e., the ability to recall a certain number of items or themes). Furthermore, only a few test versions exist which are typically administered in controlled settings (i.e., in the laboratory or clinic) in a cross-sectional manner thus precluding a fine-grained examination on a daily basis of the relationship to clinical state and treatment. In *toto* this limits scientific progress in terms of applications within psychiatry and role as a future biomarker or digital phenotype for personalized medicine purposes.[4,5] To address this, we exploited the fact that verbal recall is expressed via speech and that this data stream is potentially suited to processing with modern speech technologies. Our methodology thus moves current assessment practice towards a complete and viable process - from task presentation to automated scoring - by leveraging speech technologies for (i) the administration of the task, (ii) the transcription of voice to text, and (iii) then the application of machine learning logic from previously rated transcripts to produce automated ratings that simulate expert human ratings. This new assessment framework affords a plethora of novel opportunities of clinical value such as frequent monitoring, remote assessment of memory, and most fundamentally enables a detailed examination of the variability in memory at an individual level which can thus be a critical outcome measure for future clinical trials.[6,7,8]

We developed a series of verbal memory tests for frequent and self-administrated data collection via smart devices. In the verbal memory task, participants were asked to both immediately - and then after a delay - retell a story that was told to them via the device's loudspeaker. Ten different stories were developed (e.g., describing what happened at a birthday party) or instructions (e.g. how to assemble a skateboard) such that the stories would be different each day, and that in principle hundreds of stories could be developed to afford a more nuanced and frequent assessment of verbal memory than current tools such as the Wechsler Memory Scale[9]

and Repeatable Battery for the Assessment of Neuropsychological Status.[10] To leverage speech processing technology, the device recorded responses and we derived automated ratings on the text resulting from human transcription as well as automatic speech recognition. We expected automated ratings to correlate well with human ratings. To minimize risk of a usual scenario where machine learning methods are viewed as a mysterious "black box" as a lack of transparency and explainability can make it difficult to understand how an algorithm derived its solution,[11] we sought to keep our rating model simple and interpretable by including only a subset of possible computational features. High correlations between machine scores and human ratings would inspire confidence that employing automated methods can both complement traditional methods,[12,13] and provide a framework in which verbal memory assessment can be a core component of a system for the frequent and longitudinal monitoring of mental states.

Results

Administering verbal memory tests using smart devices

Participants (104 participants, including 25 patients with serious mental illness tested in outpatient care settings; Table 1) were able to easily understand the tasks presented and produced responses and recordings that were of sufficiently high quality such that they were suitable for analysis (Figure 1, panel A). Ninety two percent of the total of 1035 speech responses were amenable to further processing (86% for patients; Figure 1, panel B), a critically important finding given that most research on speech has been conducted in controlled laboratory settings. The retellings were on average 61 words (healthy participants' mean = 62.2 words, SD = 21.4, and patients' mean = 48.7 words, SD = 22.4; Cohen's d = -0.8, t = -9.1, $p < 0.001$), with a skew towards more short (< 10 words) responses in patients (e.g., "*I don't remember*"; healthy = 5.4%, patients 19.7% - Table 2).

Table 1 Description of participants and story recall trials

	Patients (N = 25)		Healthy (N = 79)	
	M (SD)	Range	M (SD)	Range
Age, years (SD)	49.7 (10.4)	30.0 - 67.0	21.7 (1.4)	18.0-26.0
Education, years (range)	12.3 (1.4)	7.0 - 16.0		12.0-13.0
% Female	52.2%		62.0%	
Brief Psychiatric Rating Scale*				
Affective	2.1 (1.0)	1.0-5.3		
Agitation	1.6 (0.6)	1.0-3.8		
Positive	2.2 (1.2)	1.0-5.5		
Negative	2.1 (1.0)	1.0-5.5		
Number of story recall trials	354		681	
Responses with recognizable speech** (%)	86.0%		95.9%	
Responses*** < 10 words (%)	19.7%		5.4%	

* = Presence of symptoms rated on a 1-7 scale (not present-extremely severe)

** = Words detected by human transcribers and both ASR systems

*** = Responses without recognizable speech defined as having a word count of 0

Rating the performance of recall

Human ratings of recall recordings showed the expected pattern where healthy participants received higher scores than patients. This was expected due to numerous differences between the two groups on factors such as illness, age, and education, and dictates that group differences in this study are not be interpreted as specific to memory functions *per se*. To assess the recall performance, we had expert human raters (three to seven raters) listen to each recording and rate the recall response on a 0-6 scale that we developed to capture the quality of the recall in terms of concepts and themes that were recounted (Figure 1, panel C). The average rating for recall of concept and theme was 4.3 (SD=1.3), with higher ratings assigned to responses by healthy participants (Mean = 4.6, SD = 1.1) than by patients (Mean = 3.3, SD = 1.3, Cohen's d = -1.1, t = -9.1, $p < 0.001$; Table 2). On average, each of the individual raters scores correlated with the gold standard rating at R = 0.83 (ranging between R = 0.73-0.89), and it was this level of reliability of rating that we expected an automated procedure to operate within, if it is to be considered sufficiently robust so as to be useful. Among the 21 pairs of raters, the average inter-rater correlations at the response level was 0.73, which supports the notion that the human raters were able to employ the rating scale quite reliably. As is desirable with a task design that seeks to be sensitive to differences, there was a large variance in performance, notably in patients. We conclude that the administration procedure was successful in collecting speech responses that could serve as the basis for assessing verbal memory performance.

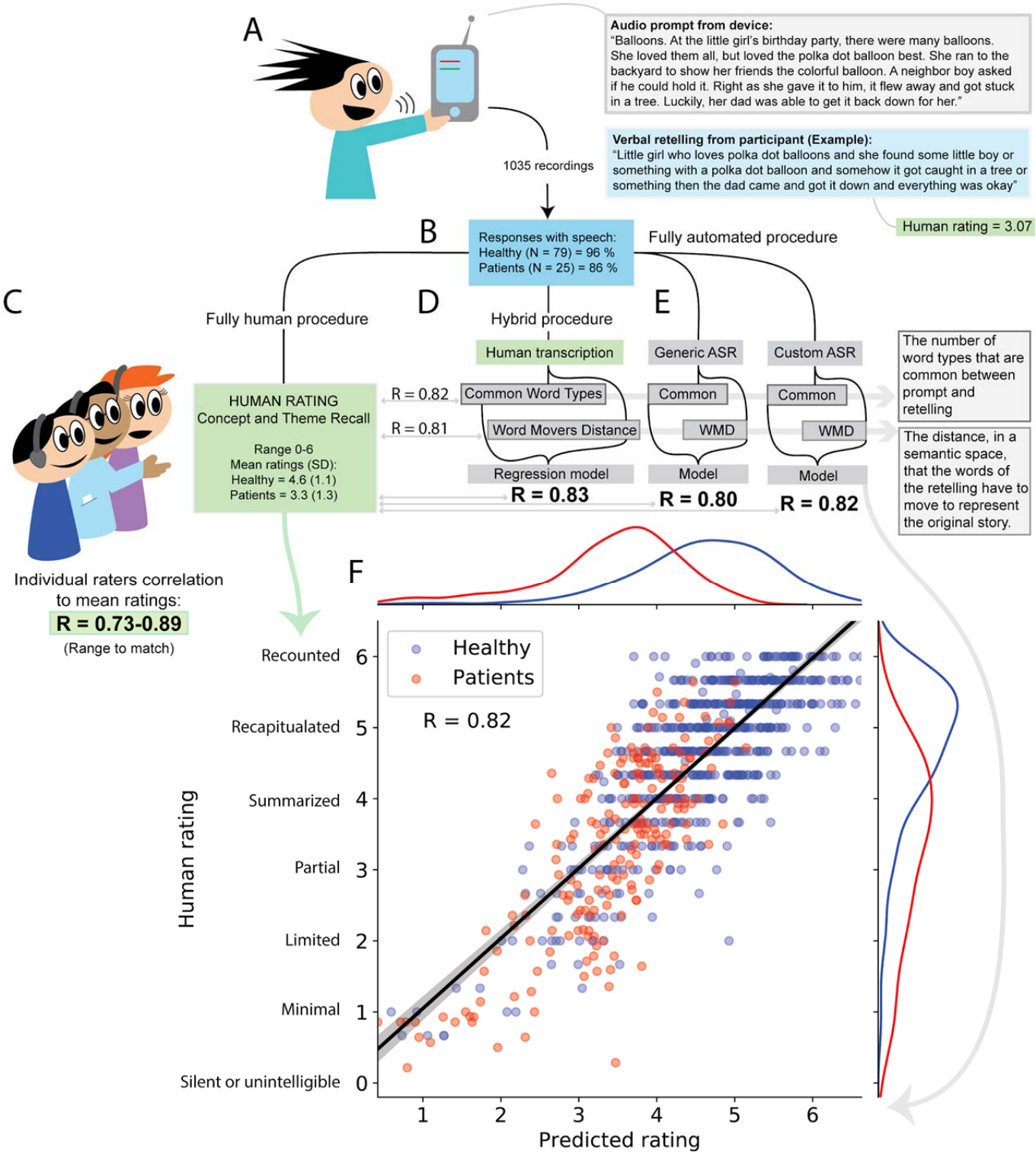


Figure 1. A summary of the procedure for administration and analysis of verbal memory using smart devices. *Panel A:* In this example, a story was presented about a girl and her balloons at a birthday party. The participants were asked to "Remember the balloon story, so you can re-tell it again later", and both immediate and delayed recall was

assessed. *Panel B:* A total of 104 participants were tested. Patients tolerated the task but had more trials where they did not provide verbal responses. *Panel C:* Humans listened to the responses and rated them for accuracy on a scale between 0 and 6. Our ground truth measure was the average of multiple ratings, and the individual raters correlated with this ground truth between R = 0.73 and R = 0.89. *Panel D:* Humans transcribed the response recordings, and the similarity of these transcriptions to the original story was compared. Two features of similarity were extracted, namely a word count procedure and a measure of distance in a semantic space. A regression model produced predicted ratings, and these correlated with average human ratings at R = 0.83, well within the range of individual human raters. *Panel E:* The same computational procedure was used on transcripts derived using generic and customized automatic speech recognition systems. The performance of the automated predictive model was still within the level of individual raters with predicted scores correlating with the average human ratings at R = 0.82. *Panel F*: A linear model based on transcriptions from the custom ASR system predicted the human ratings well, except for a tendency to assign a higher score to some short responses.

Can automated assessment methods emulate human ratings?

Automated assessment of verbal recall requires both that speech recordings are converted to text, and that there is a method to compare the resulting text to the original story in order to evaluate the amount of details remembered. To examine the viability of these different components, we first examined results generated via a procedure where humans transcribed the recordings (Hybrid procedure; Figure 1, panel D), before secondly employing *generic* ('off-the-shelf') automatic speech recognition (henceforth ASR), and then finally a *customized* ('in-house-developed') ASR (Figure 1, panel E).

Common word counts. Simply counting the number of words that were in common between the transcriptions and the original story was highly predictive of human ratings. That higher word counts generally result in higher scores is well documented in other fields (e.g., the automatic grading of essays in education[14]). The correlation between this nonlinguistic surface feature and the average human ratings was R = 0.82. This is a logical finding since the similarity will depend upon the complexity of the materials produced (i.e., the actual recall), and repeating a diverse and complete set of words should correspond to an impression of a good recall performance. Healthy participants produced more common word types (Mean = 26.7, SD = 8.1) compared to patients (Mean = 16.4, SD = 6.8, Cohen's d = 1.4, t = 17.8, $p < 0.001$; see Table 2).

Semantic similarity measures. The accuracy of the automated ratings could be further improved by including measures of utterances that were semantically similar to the original words but not identical (e.g., “father” versus “dad”) using word vector methods. Word vector methods utilize mathematical techniques where a spatial representation of a word meaning is created by analyzing the co-occurrence of words in large language corpora.[15,16,17,18] In these so-called meaning-spaces, words that co-occur and have similar meanings are located close to each other, thus allowing for the use of distance as a measure of semantically similarity. The metric Word Mover’s Distance[19] is suitable for comparing the similarity of the original story to the actual recall because it captures the meaning of words as well as a notion of how semantically distant each word in a text is to its closest aligned word in the other text on which it is to be compared. The Word Mover’s Distance between the recall and the original story correlated with the average human raters ($R = - 0.81$), and healthy participants produced recalls with shorter distances (i.e., more similar) to the original story (Mean = 1.3, SD = 0.4) compared to patients (Mean = 1.7, SD = 0.5, Cohen's $d = -1.0$, $t = -12$, $p < 0.001$).

Table 2 Description of calculated measures, by group and transcription method

	Patients (N = 25)		Healthy (N = 79)				
	Mean	SD	Mean	SD	d	t	p
Human rating (0-6)	3.3	1.3	4.6	1.1	1.1	13.4	<0.001
Word count	48.7	22.4	65.2	21.4	0.8	9.1	<0.001
Common types, calculated from:							
Human transcription	16.4	6.8	26.7	8.1	1.4	17.8	<0.001
Generic ASR	14.4	6.5	25.4	7.9	1.5	19.7	<0.001
Custom ASR	16.5	6.5	26.7	7.9	1.4	18.3	<0.001
Word Mover's Distance, calculated from:							
Human transcription	1.7	0.5	1.3	0.4	-1.0	-12.0	<0.001
Generic ASR	1.8	0.5	1.3	0.4	-1.2	-14.3	<0.001
Custom ASR	1.7	0.4	1.3	0.4	-1.1	-12.6	<0.001
Predicted scores, calculated from:							
Human transcription	3.4	0.9	4.6	0.9	1.3	15.4	<0.001
Generic ASR	3.4	0.8	4.6	0.9	1.4	17.8	<0.001
Custom ASR	3.4	0.9	4.6	0.9	1.3	15.8	<0.001

d = Cohen's d

t = t-value from Welchs t-test

p = p-value from Welchs t-test

Combined feature model. The count of common words and the semantic similarity measurements were combined in an ordinary least squares linear regression model to predict human ratings on par with individual human raters. The weighted model correlated with human ratings at R = 0.83 (range 0.74-0.90 across 5 cross-validation folds), with a regression coefficient of 0.15 for common word types, and -0.54 for Word Mover's Distance. The combined model accounted for an additional 2% of the variance over just using the simpler measure of common word types, which is not hugely impressive, but the resulting model is more robust against loss of score due to use of words that are not exactly the same as in the original story, but nonetheless have similar meaning (e.g., synonyms). The overall model provides a good fit to the average human ratings, accounting for 69% of the variance and performing at, or just slightly above, the

average human raters. Not surprisingly, computed ratings were different between groups, with retellings from healthy participants receiving higher predicted ratings (Mean = 4.6, SD = 0.9) as compared to those from patients (Mean = 3.4, SD = 0.9, Cohen's d = 1.3, t = 15.4, $p < 0.001$). This finding is as expected and simply strengthens the notion that this automated procedure to speech provides valid scores with sufficient variability that can be leveraged in future studies to detect significant cognitive changes within patients across time (i.e., sensitive enough to be used within participants). Indeed, we note that the traditional concern about 'matching' groups in a classic clinical sense is both less necessary and more improbable for machine learning studies that specifically can leverage this enormous variability that is inherent in large and 'messy' data sets.[20]

Fully automated: Speech-to-text using machines for response transcripts. To examine the viability of a fully automated system, we used two ASR systems to automatically transcribe speech and compared accuracy with human transcriptions, and found them both to be efficient and accurate. The retelling of specifically constructed and presented stories has the benefit that the participant does not have to reveal any personal or sensitive information, and we had the resources and opportunity to screen the recordings for any unprompted instances of such information, ensuring that sensitive information was not uploaded to the cloud-based ASR system. However, more effective use of cloud based tools is possible by implementing and maintaining advanced architectures for data management that can be compliant with the strictest legislations, processing data of any level of sensitivity in a safe manner.

Automatic speech recognition performed using the latest Google's speech-to-text service produced an overall word error rate of 23.3%, with lower error rates in healthy participants (17.1%) compared to patients (43.7%; see Figure 2, panel A). This high error rate is likely due to the fact that the Google language model was trained on general language rather than the language specific to our task. Even so, the predictions of a combined feature model based on transcriptions from the generic ASR procedure correlated surprisingly well with human ratings at R = 0.80 (range 0.74 - 0.88 across five folds). The robustness of such models in the context of high word error rates has been demonstrated in other domains[21] and is attributable to errors being made mostly on non-essential words, with the arguably more important common type words generally being transcribed correctly.

The word error rate using the customized ASR system was notably lower, with an overall word error rate of 10.5%. In the customized ASR system the language model was specifically

tuned towards detecting words that were likely to occur based on our stimulus material (e.g., “balloons” and “skateboards”, not “baboons” and “steakhouse”; for details on equivalent methods, see[22]). Speech from healthy participants was still detected more accurately (6.2% error rate, compared to 24.8% on speech from patients; Figure 2, panel A). Correlations between computed ratings based on transcriptions from the custom ASR procedure and the average of the human raters remained very high at 0.82 (range 0.74 - 0.88 across five folds), which was in the range of human to human agreement of 0.73 to 0.89 (Figure 1, panel G, shows the predicted ratings versus the actual human ratings based on the regression model for the automated transcripts). Importantly, the predicted ratings from fully automated procedures correlated highly with results derived using the procedure where humans transcribed the recordings (R = 0.96-0.99; Figure 2, panels B and C).

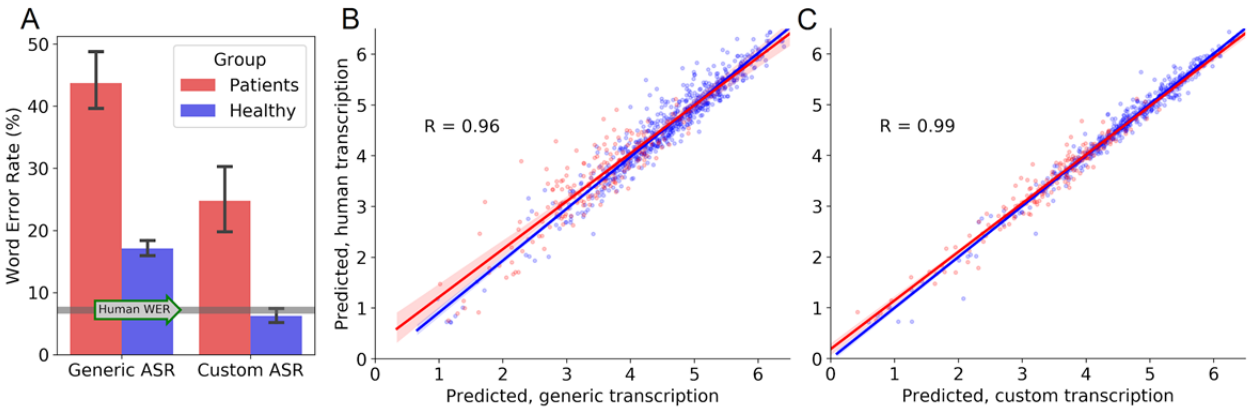


Figure 2: Accuracy of the automatic speech recognition (ASR) systems was different between the two ASR approaches and the two groups, but this did not have a large effect on predicted ratings. *Panel A:* The ASR had a lower word error rate on responses from healthy participants compared to responses from patients. The word error rate was also lower on an ASR system customized to the verbal memory task, compared to a generic, off-the-shelf system. The custom system approached the level of errors from human transcribers (7.2%), level indicated by the grey horizontal line. Error bars represent the 95% confidence intervals of the means. *Panel B:* Scores from a predictive model using natural language processing methods on human transcription was highly correlated with scores derived using transcriptions from a generic system with higher word error rates. *Panel C:* Scores derived from transcriptions using a customized ASR system with lower error rates correlated even better with scores derived using the resource-demanding human transcription procedure, arguably producing equivalent results.

In sum, the overall prediction performance derived from transcripts typed by humans (R^2 = 0.69) versus those automatically derived using ASR (R^2 = 0.67) decreased only by 2%, as expressed by variance, a value that in the current assessment context is modest and renders a fully-automated system most certainly viable and robust enough to produce data that are clinically useful.

Discussion

The current study demonstrates the viability and robustness of a method for the frequent and automated testing and scoring of verbal memory that is sufficiently robust that it can be administered outside of controlled settings and where appropriate can be self-administered by the patient themselves. Overall, the procedure was tolerated well by patients and generated high quality speech data. Natural language processing techniques were applied to the speech data and shown to provide novel ways of assessment and scoring of verbal memory. This new framework enables a detailed examination of the stability of memory and its relationship to fluctuations in clinical state within the individual, and as such may generate the critical assays for personalized medicine purposes.[4] Furthermore, this approach offers a practically viable method in ambulatory settings where mobile technology is increasingly used for clinical purposes.[23]

Beyond affording a practical tool, automated semantic techniques can measure more subtle differences in language use that cannot be seen simply in word overlap (e.g., verbatim responses). Although semantic analysis techniques have been shown previously to be effective in measuring the quality of verbal memory by going beyond counting linguistic units and/or themes and accurately measure narrative memory,[24,25,26] the current study extends previous findings notably by improving upon previously employed methods (Latent Semantic Analysis[15]) by using larger corpora and more modern semantic analysis techniques. This performance improvement is likely due to the fact that newer semantic spaces - such as the one employed in this study - use larger context vectors based on millions of words and incorporate new techniques to measure distance among vectors (e.g., Word Mover's Distance). From an assessment perspective, this information helps inform that the underlying mechanisms in recall must account for more than rote word recall, and must consider how subtle language transformations, such as recalling the gist,[27] results in accurate recalls.

Combining counts and semantic measures improved on overall prediction and can conceptually be considered a robust baseline of what computational approaches can achieve. Indeed, the 0.83 correlation to the average of the raters of the regression model using the best transcriptions is equal to the 0.83 average correlation of human raters to the average of the other raters. Multivariate models like this can have both performance and utility improved by adding more relevant features, depending on what specific aspects of assessment is important in the context. Intuitively, it may be that additionally weighting the syntax, mostly ignored by the metric

of number of common spoken word types, may improve the correlation further. A challenge with developing scoring techniques that depend on word order and syntax is that rules can be non-transferrable to other languages, thus possibly limiting scalability and generalizability of the methods. Expanding the set of features for more clinical relevance, it may be advantageous to assay evidence of language disorganization using speech graph analysis [28,29] or measures of arousal and emotional valence,[30] using acoustic parameters linked to state-dependent fluctuations in psychiatric symptoms.[31] Although we illustrate the computational natural language processing approach with a test task that is structurally similar to the prose recall subtask (the Logical Memory task) of the Wechsler Memory Scale, the techniques discussed here can likely be applied successfully to other tests with verbal responses, and more broadly to assess if a spoken utterance is relevant to its conversational context.

Transcription of the recalls via automatic speech recognition resulted in higher word error rates as compared to human transcription, but, perhaps surprisingly, the rating prediction model did not show an equivalent decrease in performance with the automated transcriptions. We have previously demonstrated that the same approach can be successfully applied to patients with affective and substance use disorders.[32] This was a group diverse in cognitive ability, with high variability in performance and transcription error rates, but the final verbal recall rating prediction model remained impervious to the non-ideal data. This is promising in terms of reproducibility of the approach in a variety of patient groups. The lack of penalty can partly be explained by the way error rates are calculated, in that variants of words (e.g., “skater”, “skateboarder”) many be counted as errors but not constitute important semantic differences, with such similarities being accounted for by the use of the semantic vector comparisons. Additionally, it may it may be related to the types of errors commonly made by ASR systems, namely errors of inserting, deleting or substituting short and frequently used words like “is”, “in” and “the”, as well as filled pauses such as “uh”, words that will have less consequence when assessing recall performance.[33] Errors may also be specific to certain disorders or accents. Shor et al. found that the five most mistaken phonemes accounted for 20% of errors in a sample of patients with amyotrophic lateral sclerosis, underlining the potential for specialized and tailored speech recognition models for applications in medical settings.[34]

Future studies need to validate the current findings in terms of both the sensitivity to detect changes in memory over time within individuals and establish whether such changes are clinically

meaningful at an individual level.[35] Prior work that has employed word vector methods to characterize language in psychiatric patients in terms of semantic coherence has found it useful in predicting differences in patients with schizophrenia and risk of illness onset in psychosis,[36,37,38,39,40,41] but also that such approaches provide new metrics for analysis of performance in well-established neuropsychological tests.[42] Although our current study does not explicitly establish whether such frequent monitoring is psychometrically viable for ambulatory purposes, we can extrapolate that the current design enables frequent monitoring and this specific task can be administered by smart devices both within clinical settings and remotely. Naturally, conducting mental state assessments outside of the controlled setting comes with several practical, technical and legal challenges.[43] Nonetheless, for those patients who have access to digital devices, and can operate such devices with minimal supervision, future assessment methods that embrace mobile technologies promise to be of enormous value in psychiatry and may even enhance the bond between patients and clinicians.[44]

Methods

Participants

The participant sample comprised 104 adults. Twenty-five patients were recruited from a group home facility in the Southeastern US (Mean age = 49.7 years; SD = 10.4 years, 52.2% female), all met U.S. federal definitions of serious mental illness (per the Alcohol, Drug Abuse and Mental Health Services Administration Reorganization Act[45]) and were receiving treatment from a multidisciplinary team. Of these, two-thirds met the criteria for schizophrenia (N = 16), and the remaining major depressive disorder (N = 8) and bipolar disorder (N = 1), as established after structured clinical interviews (Structured Clinical Interview for DSM–IV–TR[46]). The severity of illness in patients was assessed using the Brief Psychiatric Rating Scale[47] (Table 1). In this scale the severity of self-reported symptoms and observed signs are rated on a scale of 1 (not present) to 7 (extremely severe), and items (e.g., hallucinations, excitement) are combined into "Affective", "Agitation", "Positive" and "Negative" symptom categories. The categories were based on a factor solution[48] with some minor modifications to attain acceptable internal consistency, and diagnoses and symptom ratings reflected consensus from the research team. The average scores presented in Table 1 indicated that the sample was relatively asymptomatic overall at the time of testing but there was considerable variability with cases of reported moderate and above severity, represented by cases having category average scores of up to 5.5. Such averages can hide elevations on particular items (e.g., a patient with extreme values within the "Agitations" category may still have an average score of 3.8). The other participants (N=79) were undergraduate students at Louisiana State University presumed to be healthy (henceforth termed 'healthy participants'; mean age = 21.7 years; SD = 1.4 years, 62% female). The research program was approved by the relevant ethics committee (LSU Institutional Review Board #3618) and all participants provided their informed written consent.

Procedure and Materials

The recall tasks were developed to run on an iOS software environment - a mobile operating system created and developed by Apple Inc. - and were a part of a larger set of assessment tasks that engaged participants in spoken and touch-based interactions to capture structured daily measures of cognition, motor skills, and language.[29,41] Ten text passages were

developed that were to be remembered and retold by the participants. Five of the passages were narrative stories and five of the texts were instructions on how to perform certain actions. The narrative stories were structurally similar to the Logical Memory subtest of the widely used Wechsler Memory Scale[8] and were between 69 and 87 words in length (average length = 75 words; see Figure 1 and Supplementary material for examples). Each narrative had two characters, a setting, an action that happened in the setting causing a problem, and then a resolution. The instructional passages started with a statement or question about an action that was to be performed, continued with description on how to accomplish the goal of the action, then ended with some concluding details (62 and 83 words in length, average length = 73 words). The passages were presented orally in a male voice and the participant was asked to retell the story immediately with as many details as possible, as well as a second retelling of the same prompt later in the testing session. The mobile device recorded the participant's retelling.

Every response recording was rated for accuracy on a 0-6 scale by human raters with clinical experience, and the details of the rating rubrics are in the Supplementary Material. The average of these ratings was treated as the gold standard that the automated modelling approach was designed to predict.

Each recall was independently transcribed by two human transcribers and differences in transcription were resolved, producing an overall human word error rate of 7.2%. Although highly unlikely given the neutral nature of the story recall task, in the rating and transcription procedures the recordings were checked for the presence of directly identifiable information (e.g., names) or sensitive health information. Machine transcription was conducted on the pre-checked recordings using Google's speech-to-text transcription (https://cloud.google.com/speech-to-text/). However, since such generic tools are built to accommodate speech on a wide variety of topics we also built a customized speech recognizer based on the Kaldi speech recognition toolkit.[49] See Supplementary Material and [19,30] for further details on the transcription procedures.

Natural Language Processing features for automated rating of passages

When scoring a recall, human raters compare the similarity of the actual recall to the original story that the participant was presented with. In order to create an automated way to score recall it is therefore necessary to develop a model that simulates this process, albeit based upon transcriptions rather than audio *per se*. Therefore we selected linguistic surface features (e.g., word counts) and semantic content features of the transcribed recall responses to derive a composite score to compare with the human ratings. Prior to analysis, the transcriptions were 'preprocessed' (using the built-in string processing methods in the Python programming language; Python Software Foundation, https://www.python.org/) to render suitable for computational methods by for example transforming all text to lowercase, removing punctuation (e.g., commas, periods) and instances of transcribed hesitation markers (e.g., "uh").

Common word type count and semantic similarity measures. First, we computed a simple surface feature describing the similarity between prompt and recall, namely the raw counts of the number of occurrences of particular word types (i.e., individual words only counted once) that were in common between the response and the original prompt.

Second, we computed the distance between the prompt and the recall in a semantic vector space. This means that the words in the original prompt and also the participants' recall were converted to numerical vector representations that convey the semantic content of the recall. With this method, words are "embedded" in a multidimensional space, where the vectors represent the locating coordinates for words in a way that words with similar meanings are located closer together. These spaces are derived by means of computational language models that are based on analyzing the co-occurrence of words in large language corpora.[15,16,17,18] We utilized a set of publicly available word embeddings based on a semantic space with 300 dimensions derived from training a Word2vec model on 240 million words from the Google News corpus.[17] Critically, the semantic similarity between two words vectors may be calculated by measuring "distances" in semantic space between words in the response and their closest related words in the original story presented. Even when the discourse from the prompt and recall have no words in common, based on the embedded word vectors that capture aspects of the semantics, the metrics can assess the "distance" between the two stories (i.e., prompt and recall) in a meaningful way. We calculated this distance between the recall and the original story with the Word Mover's Distance metric[19]

(using the Gensim software package.[50]). Such metrics should produce a measure of the amount of semantic information in common between the recall and the original text.

Regression model for predicting ratings. The two similarity measurements were used as independent variables in an ordinary least squares regression model to estimate the human scores. The correlation between the estimated values and the average human rating was our main performance metric, and to minimize bias in our assessment of model performance we estimated the coefficient using a 5-fold cross validation procedure. This procedure involves dividing the data into 5 subsets, building the linear model on four of the subsets (i.e., the training sets) while leaving one subset (i.e., the test set) for estimating the correlation coefficient, and repeating this procedure in a fashion such that all subsets had served as both training and test sets. Both the linear models and the cross validation procedure were implemented using the scikit-learn Python module.[51]

If the predicted scores correlated well with human scores, we may justifiably employ such an automated metric to measure the fidelity of the recall responses with reference to the original story presented. This kind of performance metric would thus be automated, consistent and objective.

Funding

This work was supported by the Research Council of Norway (grant number 231395 to Brita Elvevåg).

Acknowledgements

We thank Taylor L. Fedechko, Tovah M. Cowan and Thanh P. Le for their assistance during parts of the project. The authors declare that there are no conflict of interests.

References

1. Aleman, A., Hijman, R., de Haan, E. H. F. & Kahn, R. S. Memory impairment in schizophrenia: a meta-analysis. *Am. J. Psychiatry.* **156,** 1358–1366 (1999). doi:10.1176/ajp.156.9.1358
2. Cirillo, M. A. & Seidman, L. J. Verbal declarative memory dysfunction in schizophrenia: from clinical assessment to genetics and brain mechanisms. *Neuropsychol. Rev.* **13,** 43-77 (2003). doi:10.1023/A:102387082
3. Skelley, S. L., Goldberg, T. E., Egan, M. F., Weinberger, D. R. & Gold J.M. Verbal and visual memory: characterizing the clinical and intermediate phenotype in schizophrenia. *Schizophr. Res.* **105,** 78-85 (2008). doi:10.1016/j.schres.2008.05.027
4. Insel, T. R. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA* **318,** 1215-1216 (2017). doi:10.1001/jama.2017.11295
5. Hsin, H., Fromer, M., Peterson, B., et al. Transforming Psychiatry into Data-Driven Medicine with Digital Measurement Tools. *NPJ Digit. Med.* **1** (2018). doi: 10.1038/s41746-018-0046-0
6. Bucci, S. et al. Actissist: Proof-of-Concept Trial of a Theory-Driven Digital Intervention for Psychosis. *Schizophr. Bull.* **44,** 1070-1080 (2018). doi:10.1093/schbul/sby032
7. Schlosser, D. et al. Efficacy of PRIME, a Mobile App Intervention Designed to Improve Motivation in Young People With Schizophrenia. *Schizophr. Bull.* **44,** 1010-1020 (2018). doi:10.1093/schbul/sby078
8. Stroud, C., Onnela, J-P. & Manji, H. Harnessing digital technology to predict, diagnose, monitor, and develop treatments for brain disorders. *NPJ Digit. Med.* **2,** 1-4 (2019). doi: 10.1038/s41746-019-0123-z
9. Wechsler, D. *Wechsler Memory Scale - Third Edition, WMS-III: Administration and scoring manual.* (The Psychological Corporation, San Antonio, TX, 1997).
10. Randolph, C., Tierney, M., Mohr, E. & Chase, T. The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): Preliminary Clinical Validity. *J. Clin. Exp. Neuropsychol.* **20,** 310-319 (1998). doi:10.1076/jcen.20.3.310.823
11. Tandon, N. & Tandon, R. Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia, *Schizophr. Bull.* **44,** 939–941 (2018). doi:10.1093/schbul/sby101

12. Lehr, M., Prud'hommeaux, E., Shafran, I. & Roark B. Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In Proceedings Interspeech, Portland, OR, USA,1039-1042 (2012). URL: https://www.isca-speech.org/archive/interspeech_2012/i12_1039.html

13. Lehr, M., Shafran, I., Prud'hommeaux, E. & Roark, B. Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, 211-220 (2012). URL: https://www.aclweb.org/anthology/N13-1021

14. Foltz, P. W., Streeter, L. A., Lochbaum, K. E. & Landauer, T. K. Implementation and applications of the Intelligent Essay Assessor. In Shermis M., Burstein J, eds. *Handbook of Automated Essay Evaluation* (pp. 68-88, Routledge, New York, NY, 2013).

15. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* **104,** 211-240 (1997). doi:10.1037/0033-295X.104.2.211

16. Bengio, Y. et al. A neural probabilistic language model. *Journal of Machine Learning Research* **3,** 1137-1155 (2003). doi:10.1162/153244303322533223

17. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. In: *Workshop Proceedings for International Conference on Learning Representations 2013*. https://arxiv.org/abs/1301.3781 (2013).

18. Pennington, J., Socher, R. & Manning, C. D. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 1532-1543 (2014).

19. Kusner, M., Sun, Y., Kolkin, N. & Weinberger, K. From Word Embeddings To Document Distances. *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 957-966 (2015).

20. Chandler, C., Foltz, P.W. & Elvevåg, B. Using machine learning in psychiatry: The need to establish a framework that nurtures trustworthiness. *Schizophr. Bull.* (in press).

21. Foltz, P. W., Laham, D. & Derr, M. Automated Speech Recognition for Modeling Team Performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Santa Monica, CA, USA, 673–677 (2003). doi:10.1177/154193120304700402

22. Cheng, J. Real-time scoring of an oral reading assessment on mobile devices. In: *Proceedings Interspeech*, Hyderabad, India, 1621-1625 (2018). doi:10.21437/Interspeech.2018-34

23. Carlo, A. D., Hosseini G. R., Renn, B. N., & Areán, P. A. By the numbers: ratings and utilization of behavioral health mobile applications. *NPJ Digit. Med*. **2** (2019). doi: 10.1038/s41746-019-0129-6

24. Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A. & Flicker, L. Latent semantic analysis: A new method to measure prose recall. *J. Clin Exp. Neuropsychol.* **24,** 26-35 (2002). doi:10.1076/jcen.24.1.26.965

25. Lautenschlager, N. T., Dunn, J. C., Bonney, K., Flicker, L. & Almeida, O. P. Latent semantic analysis: an improved method to measure cognitive performance in subjects of non-English speaking background. *J. Clin. Exp. Neuropsychol.* **28,** 1381-387 (2006). doi:10.1080/13803390500409617

26. Rosenstein, M., Diaz-Asper, C., Foltz, P. W. & Elvevåg, B. A computational language approach to modeling prose recall in schizophrenia. *Cortex* **55,** 148-166 (2014). doi:10.1016/j.cortex.2014.01.021

27. Kintsch, W. The role of knowledge in discourse comprehension: a construction-integration model. *Psychol. Rev.* **95,** 163–182 (1988). doi:10.1037/0033-295X.95.2.163

28. Mota, N. B., Copelli, M. & Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia* **3,** 1-10 (2017). doi:10.1038/s41537-017-0019-3

29. Cabana, A., Valle-Lisboa, J., Elvevåg, B. & Mizraji, E. Detecting order-disorder transitions in discourse: Implications for schizophrenia. *Schizophr. Res.* **131,** 157-164 (2011). doi:10.1016/j.schres.2011.04.026

30. Cheng, J., Bernstein, J. & Rosenfeld, E., et al. Modeling self-reported and observed affect from speech. In *Proceedings Interspeech*, Hyderabad, India, 3653-3657 (2018). doi:10.21437/Interspeech.2018-2222

31. Cohen, A. S. et al. Ambulatory vocal acoustics, temporal dynamics, and serious mental illness. *J. Abnorm. Psychol.* **128,** 97-105 (2019). doi:10.1037/abn0000397

32. Chandler, C. et al. 2019. Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Minneapolis, Minnesota, USA, 137–147 (2019). URL: https://www.aclweb.org/anthology/W19-3016

33. Stolcke, A. & Droppo, J. Comparing Human and Machine Errors in Conversational Speech Transcription. In: *Proceedings Interspeech* 2017, 137-141 (2017).
34. Shor et al. Personalizing ASR for Dysarthric and Accented Speech with Limited Data. https://arxiv.org/abs/1907.13511 (2019)
35. Torous, J., Staples, P., Barnett, I., Onnela, J. P. & Keshavan M. A crossroad for validating digital tools in schizophrenia and mental health. *npj Schizophrenia* **4,** 6 (2018). doi:10.1038/s41537-018-0048-6
36. Bedi, G. et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia* **1,** 15030 (2015). doi:10.1038/npjschz.2015.30
37. Corcoran, C. M. et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* **17,** 67-75 (2018). doi:10.1002/wps.20491
38. Elvevåg, B., Foltz, P. W., Weinberger, D. R. & Goldberg, T. E. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophr. Res.* **93**, 304-316 (2007). doi:10.1016/j.schres.2007.03.001
39. Elvevåg, B., Foltz, P. F., Rosenstein, M. & DeLisi L. E. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of Neurolinguistics* **23,** 270-284 (2010). doi:10.1016/j.jneuroling.2009.05.002
40. Rosenstein, M., Foltz, P. W., DeLisi, L. E. & Elvevåg B. Language as a biomarker in those at high-risk for psychosis. *Schizophr. Res.* **165,** 249-250 (2015). doi:10.1016/j.schres.2015.04.023
41. Iter, D.,Yoon, J. & Jurafsky, D. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology*, New Orleans, LA, USA, 136-146 (2018). doi:10.18653/v1/W18-0615
42. Holmlund, T. B., Cheng, J., Foltz, P. W., Cohen, A. S. & Elvevåg B. Updating verbal fluency analysis for the 21st century: Applications for psychiatry. *Psychiatry. Res.* **273,** 767-769 (2019). doi:10.1016/j.psychres.2019.02.014
43. Holmlund, T. B. et al. Moving psychological assessment out of the controlled laboratory setting : Practical challenges. *Psychol. Assess.* **31,** 292-303 (2019). doi:10.1037/pas0000647

44. Noel, V.A., Carpenter-Song, E., Acquilano, S.C., Torous, J., & Drake, R.E. The technology specialist: a 21st century support role in clinical care. *NPJ Digit. Med.* **2** (2019). doi: 10.1038/s41746-019-0137-6

45. The Alcohol, Drug Abuse and Mental Health Services Administration Reorganization Act (ADAMHA) of 1992 PL 102-321.

46. First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. *Structured Clinical Interview for DSM–IV–TR Axis I Disorders–Patient edition*. (New York State Psychiatric Institute, New York, 2002).

47. Lukoff, D., Nuechterlein, H. & Ventura, J. Manual for the expanded brief psychiatric rating scale. *Schizophr. Bull.* **12,** 594-602 (1986).

48. Kopelowicz, A., Ventura, J., Liberman, R. P., & Mintz, J. . Consistency of Brief Psychiatric Rating Scale factor structure across a broad spectrum of schizophrenia patients. *Psychopathology* **41,** 77-84 (2008). doi: 10.1159/000111551

49. Povey, D. et al. The KALDI speech recognition toolkit. In *Proceedings IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Hawaii, USA, **4** (2011).

50. Řehůřek, R. & Sojka, P. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Framework*. Valletta, Malta, 45-50 (2010).

51. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal Of Machine Learning Research* **12,** 2825-2830 (2011).

Supplementary Material

to

Applying speech technologies to assess verbal memory in patients with serious mental illness

Details on the recall task procedure

Each participant was presented with one narrative and one instructional passage per testing session with the mobile device. The presentation of stimuli items (i.e. the different passages) was counterbalanced using a rotating design such that no one received a repetition of a story across sessions and all stories were sampled across participants.

A narrative story example:

> *"Balloons. At the little girl's birthday party, there were many balloons. She loved them all, but loved the polka dot balloon best. She ran to the backyard to show her friends the colorful balloon. A neighbor boy asked if he could hold it. Right as she gave it to him, it flew away and got stuck in a tree. Luckily, her dad was able to get it back down for her."*

An instructional passage example:

> *"Skateboards. How to put your skateboard together? Set all the parts and your tools out in front of you. Start by attaching the trucks, but don't screw the front one all the way in. The trucks need to be able to move a little so you can turn. After that, attach all four wheels. Turn it right side up and test it out; then make any adjustments you need to on the front truck."*

In the case of the 5 narrative stories, they were additionally prompted to retell the story later on in the testing session (e.g. "*Retell the balloon story again now. Put in all the details you*

can remember.") an average of 17 minutes after the original audio appeared and the delayed retelling was done. The participant was given a maximum of one minute to speak, and the time remaining was indicated by a timer bar on devices' screen.

Concept and Theme Recall Rating

Trained human raters listened to the audio recordings of the recalls and assigned scores as to the quality of the *narrative concepts and recall theme* (i.e., characters, actions, feelings, motivations, names, dates, descriptors, plans, causes, situations). Scores were assigned on a 0 to 6 scale, such that zero represented "silent or unintelligible", and a high score (6) indicated that all major and almost all minor concepts and/or themes were recalled and that all facts corresponded to the original. All responses were rated by multiple raters (a minimum of 3, maximum of 7) and the average rating was computed.

Raters were given the following instructions:

> Please rate the accuracy of the recall of specific themes and concepts from the passage. These can include characters, actions, feelings, motivations, names, dates, descriptors, plans, causes, and situations that were mentioned in the passage. Small transformations in wording on concepts (e.g., "a man" "a gentleman", "a guy") can be counted as equivalent concepts. However, changes in the amount of detail (e.g., "A rainy afternoon", "An afternoon") should be considered as having a different number of concepts.
>
> 0 *Silent or unintelligible*.
> 1 *Minimal*. Few accurate concepts/themes from the original passage, with or without off-topic material.
> 2 *Limited*. Some accurate concepts/themes, but none of the major concepts/themes. Less than a third of the concepts conveyed.
> 3 *Partial*. At least half the original material missing; a few major concepts/themes are included.
> 4 *Summarized*. At least two thirds of the concepts/themes are included and accurate, including two or more important concepts/themes.
> 5 *Recapitulated*. Most major concepts/themes are included, along with many incidental concepts - accurate concepts.
> 6 *Recounted*. All major and almost all minor concepts/themes are included. All facts correspond to the original.

Transcription

Human transcription was conducted via an in-house designed web interface that allowed for audio playback, transcription input and coding of different types of noise and non-speech events and sounds. This careful process allowed us to verify that there were no explicit reference to health information, names, addresses or other possibly sensitive or personal identifying items in the recordings beyond the voice itself. Human transcriptions of speech collected with the mobile application were used to build this custom language model that was particularly designed to detect words and phrases relevant to the audio prompts. The acoustic model used for the custom speech recognition was a Deep Neural Network - Hidden Markov Model [1] trained on all training sets of the Librispeech data.[2] Further details on the speech recognition systems can be found in [3] and [4]. Word error rates for the machine transcriptions were calculated separately for patients and healthy participants (by estimating the minimal edit distance with the Wagner-Fischer algorithm using the "jiwer" software package for Python - https://github.com/jitsi/asr-wer/).

References in Supplementary Material

1. Zhang X, Trmal J, Povey D, Khudanpur S. Improving deep neural network acoustic models using generalized maxout networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy: 2014; 215-219. doi:10.1109/ICASSP.2014.6853589

2. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2015;5206-5210. doi:10.1109/ICASSP.2015.7178964

3. Cheng J. Real-time scoring of an oral reading assessment on mobile devices. In: *Proceedings Interspeech*, Hyderabad, India: 2018;1621-1625. doi:10.21437/Interspeech.2018-34

4. Chandler C, Foltz PW, Cheng J, et al. 2019. Overcoming the bottleneck in traditional assessments of verbal memory: Modeling human ratings and classifying clinical group membership. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Minneapolis, Minnesota, USA: 2019;137–147. URL: https://www.aclweb.org/anthology/W19-3016

Appendices.

1: Written instructions for the dMSE

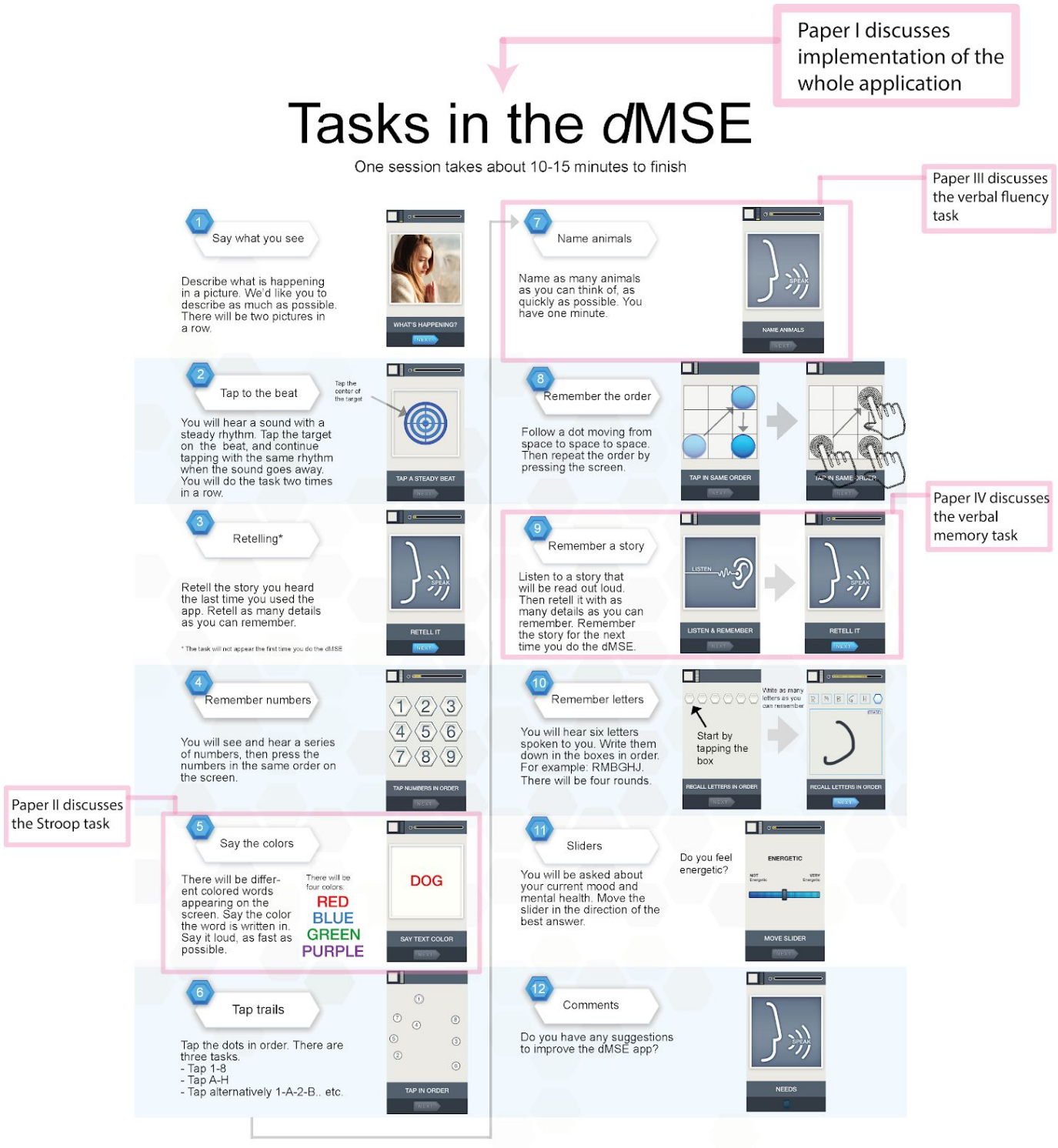
2: Online questionnaire for MinTest participants

3: Consent form - dMSE

4: Consent form - MinTest

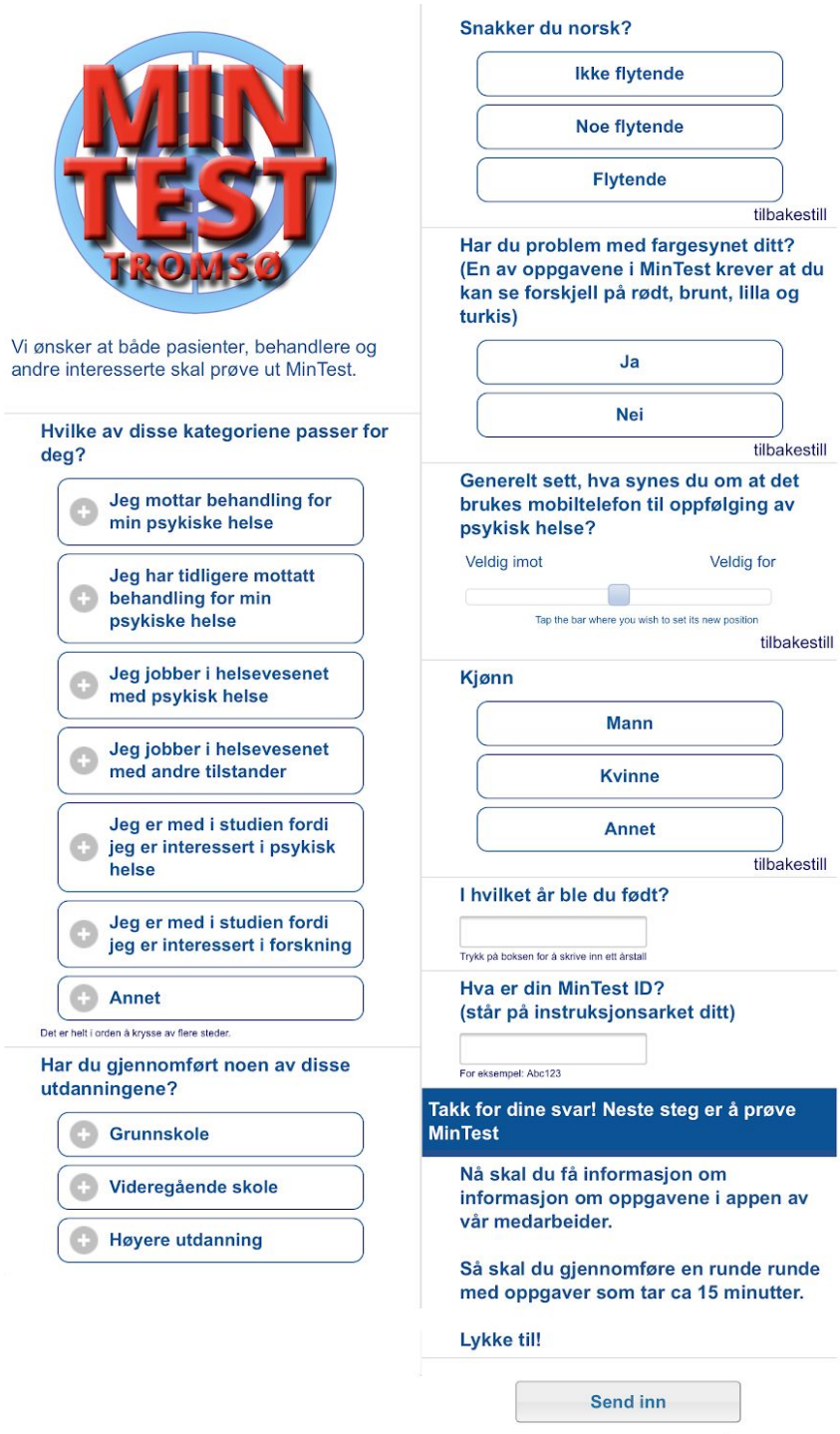
Appendix 1

An overview of the tasks in the *d*MSE and *MinTest* mobile applications with written instructions. A Norwegian version can (at the time of writing, October -19) be found at https://uit.no/mintest . The three tasks with speech responses that were selected for analysis in this thesis are highlighted with pink boxes.

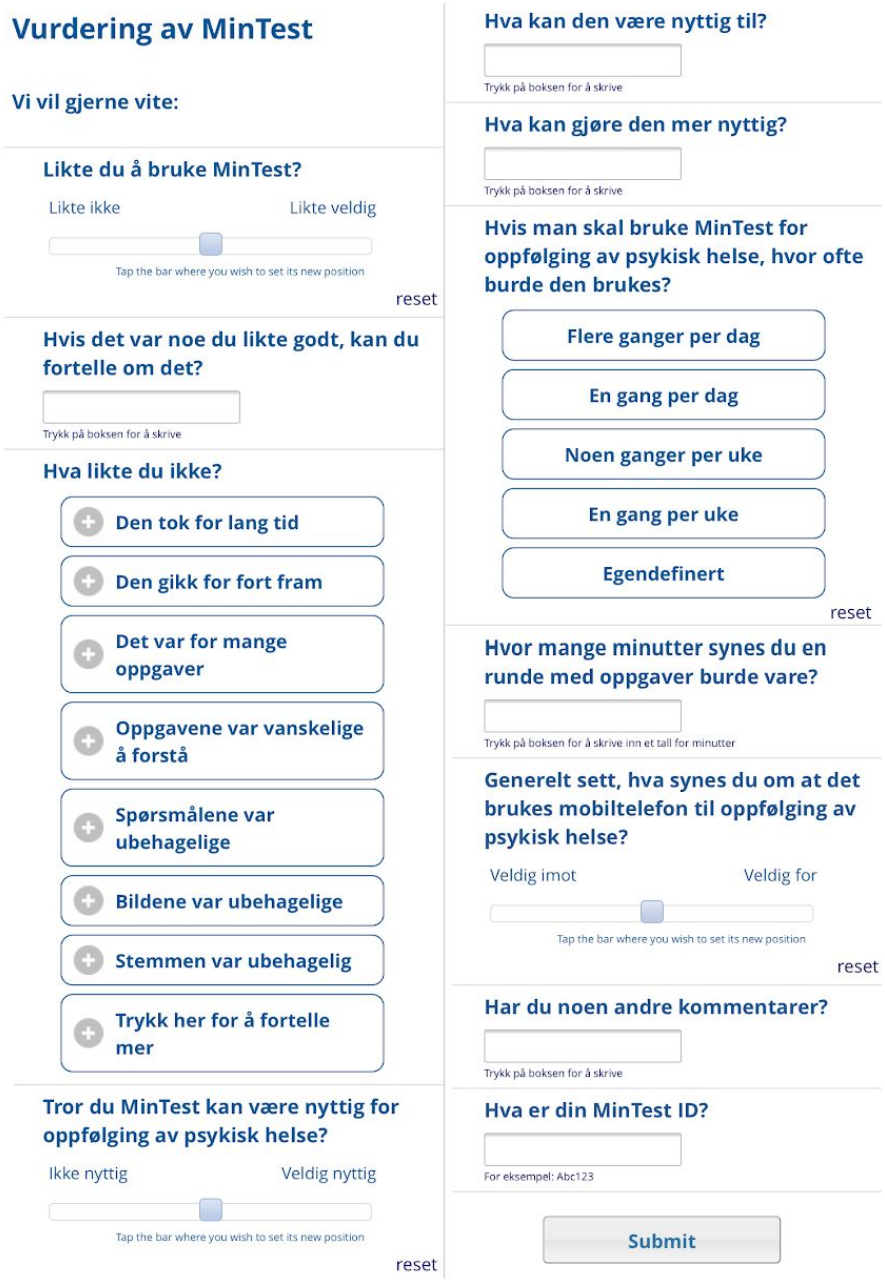


Appendix 2

Questionnaire used in Norway, before testing:



Questionnaire used in Norway, after testing:



Appendix 3

CONSENT FORM - Outpatients

Project Title: Development of Mobile Status Exam for Psychiatric Symptoms - Outpatient

Performance Site:

1. Baton Rouge Mental Health Clinic, Baton Rouge, LA
2. Tyler Mental Health Clinic, Lafayette, LA
3. Louisiana State University, Baton Rouge, LA
4. Medical Management Options, Baton Rouge, LA
5. Subjects homes, as needed.

Investigator: The following investigator is available for questions Monday-Friday, 9:00 a.m.- 4:30 p.m.

Alex S. Cohen, Ph.D., Psychology Department, LSU. (225) 578-7017

This is a consent form for research participation. It contains important information about this study and what to expect if you decide to participate. Please consider the information carefully. Feel free to discuss the study with your friends and family and to ask questions before making your decision whether or not to participate.

Purpose of the Study: The purpose of this research project is to develop a measure of mental illness symptoms using a phone. This new measure can be used to help support people with mental illness and help symptoms from getting worse and prevent hospitalization.

Inclusion Criteria: You are being asked to participate in this study because you are between the ages of 18 and 60, and are either:

1. A patient with a mental illness diagnosis (e.g., schizophrenia, schizoaffective disorder, depression, bipolar disorder, personality disorder) and are being treated by a mental health professional.
2. An individual who is free from mental illness.

Exclusion Criteria: Participation is excluded for individuals who not judged to be clinically stable.

Maximum Number of Subjects: The maximum number of subjects will be 300.

Study Procedures/Description of the Study: I am aware that this study will take place over one week. There will be two phases to this study.

The first phase will take place on two separate days with appointments each lasting approximately two hours. During these sessions, I will be asked questions about my history and about my mental illness. I will also be asked to complete questionnaires and paper and pencil

tests that measure personality, attention and memory and depression. During parts of this study, I will be recorded using a laptop computer and a microphone. For participating in these sessions, I will be compensated $20 cash for each session, for a total of $40.

During the second phase of the study, I will be given a phone to carry around with me for seven days. Once per day, I will be asked to use an app on the phone that takes approximately 20 minutes. During this time, I will play some games, I will report on my mood and symptoms, and will talk on the phone. I will be asked to talk about my mood, my thoughts and my symptoms. I will be audio and video recorded during this time – though the phone will indicate when I am being recorded. At no time will I be recorded without my permission. I will be compensated $10 for each time I use the app; for a total of $70 during the week.

The researchers would like permission to access my medical records in order to document my diagnoses and prior hospitalizations. I have the option of either giving or not giving the researchers the right to access my medical records, depending on my comfort level. There will be no penalty, reduction in compensation, or other issue for my decision either way.

Benefits: I understand that I will not directly benefit from participating in this study. My participation will help researchers develop new tools for measuring mental illness.

Risks/Discomforts: This study may be inconvenient in that it will take some of my time. I also recognize that I will be asked to talk about my mental health history, and that I will be recorded during some parts of this study. These recordings will be uploaded to a central computer for analysis. At no time will these recordings be shared with anyone not involved with the study. These recordings will be destroyed at the end of the study.

Right to Refuse: Participation in this study is voluntary. I may refuse to answer any questions or discontinue any test I am taking. Further, I can change my mind and withdraw from this study at any time without risking my relationship with Louisiana State University or any group homes or Mental Health clinics. I also recognize that I can contact the researchers at any point after the study is complete to have my audio and video taped records destroyed.

Privacy: All information obtained in this study will be kept confidential. That means my information will not be shared with anyone, unless legally compelled. Limits to confidentiality include situations where an individual is at risk of hurting themselves (e.g., suicide) or hurting someone else (e.g., homicide, child abuse). I understand that the investigators are required by law to report any reasonable suspicions.

My records will be kept in a locked laboratory in a secure facility. Electronic data will be entered without identifying information and will be password protected. To ensure confidentiality, I will be assigned a number. All information collected during this study will be linked to this number and kept separate from any identifying information such as my name. Results of the study may be published, but no names or identifying information will be included for publication.

The researchers are applying for a Certificate of Confidentiality from the National Institute of Health (NIH). This Certificate will protect the investigators from being forced to release any research data in which I am identified, even under court order or subpoena, without my written consent. This protection does not affect the investigators' legal responsibility to report information about suspected or known sexual or physical abuse of a child or about your expression of a clear and present danger of harming yourself or others to proper authorities. The Certificate does not prevent me or a member of your family from voluntarily releasing information about myself or my involvement in this study.

Financial Information: I will receive $20 cash for session one and $20 for session two. I will receive $10 for each day I complete my phone app, for a total of $70. The total compensation for this project will not exceed $110.

Withdrawal: Participation in this study is voluntary. I may withdraw from this study at any time without penalty or loss of any benefit to which I would otherwise be entitled to.

Signatures:
The study has been discussed with me and all my questions have been answered. I may direct additional questions regarding study specifics to the investigators. If I have questions about subjects' rights or other concerns, I can contact Dennis Landin, Ph.D., Chairman, LSU Institutional Review Board, (225)578-8692. I agree to participate in the study described above and acknowledge the researchers' obligation to provide me with a copy of this consent form if signed by me.

_____ _____
Participant Signature Date

I give _____ or do not give _____ permission for the researchers to access my medical records.

_____ _____
Participant Signature Date

**Research Assistant: please indicate whether the consent form was read to the participant.*

(Check One)

_____ I certify that I have read this consent form to the participant and explained that by completing the signature line above, he/she has agreed to participate (*NOTE – Consent form should be read to all patient participants)*.

_____ The participant will be enrolled as a control and is English-literate. The participant refused my offering to read this consent form to them.

_____ _____

Signature of Research Assistant Date

_____ _____

Signature of Principal Investigator Date

CONSENT FORM – Healthy Adults

Project Title: Development of Mobile Status Exam for Psychiatric Symptoms - Outpatient

Performance Site:
Louisiana State University, Baton Rouge, LA

Investigator: The following investigator is available for questions Monday-Friday, 9:00 a.m.- 4:30 p.m.

Alex S. Cohen, Ph.D., Psychology Department, LSU. (225) 578-7017

This is a consent form for research participation. It contains important information about this study and what to expect if you decide to participate. Please consider the information carefully. Feel free to discuss the study with your friends and family and to ask questions before making your decision whether or not to participate.

Purpose of the Study: The purpose of this research project is to develop a measure of mental illness symptoms using a phone. This new measure can be used to help support people with mental illness and help symptoms from getting worse and prevent hospitalization. You are being asked to complete this app to help validate and test it.

Inclusion Criteria: You are being asked to participate in this study because you are between the ages of 18 and 60, and are an individual who is free from mental illness.

Exclusion Criteria: Participation is reserved for individuals 18 years and older

Maximum Number of Subjects: The maximum number of subjects will be 300.

Study Procedures/Description of the Study: I am aware that this study will take place over one week. There will be two phases to this study.

The first phase will take place on two separate days with appointments, the first of which will last approximately two hours and the second lasting 30 minutes. During these sessions, I will be asked questions about my history and about my mental illness. I will also be asked to complete questionnaires and paper and pencil tests that measure personality, attention and memory and depression. During parts of this study, I will be recorded using a laptop computer and a

microphone. For participating in these sessions, I will be compensated $20 cash for the first session and $10 for the second, for a total of $30.

During the second phase of the study, I will be given a phone to carry around with me for seven days. Once per day, I will be asked to use an app on the phone that takes approximately 20 minutes. During this time, I will play some games, I will report on my mood and symptoms, and will talk on the phone. I will be asked to talk about my mood, my thoughts and my symptoms. I will be audio and video recorded during this time – though the phone will indicate when I am being recorded. At no time will I be recorded without my permission. I will be compensated $10 for each time I use the app; for a total of $70 during the week.

Benefits: I understand that I will not directly benefit from participating in this study. My participation will help researchers develop new tools for measuring mental illness.

Risks/Discomforts: This study may be inconvenient in that it will take some of my time. I also recognize that I will be asked to talk about my mental health history, and that I will be recorded during some parts of this study. These recordings will be uploaded to a central computer for analysis. At no time will these recordings be shared with anyone not involved with the study. These recordings will be destroyed at the end of the study.

Right to Refuse: Participation in this study is voluntary. I may refuse to answer any questions or discontinue any test I am taking. Further, I can change my mind and withdraw from this study at any time without risking my relationship with Louisiana State University or any group homes or Mental Health clinics. I also recognize that I can contact the researchers at any point after the study is complete to have my audio and video taped records destroyed.

Privacy: All information obtained in this study will be kept confidential. That means my information will not be shared with anyone, unless legally compelled. Limits to confidentiality include situations where an individual is at risk of hurting themselves (e.g., suicide) or hurting someone else (e.g., homicide, child abuse). I understand that the investigators are required by law to report any reasonable suspicions.

My records will be kept in a locked laboratory in a secure facility. Electronic data will be entered without identifying information and will be password protected. To ensure confidentiality, I will be assigned a number. All information collected during this study will be linked to this number and kept separate from any identifying information such as my name. Results of the study may be published, but no names or identifying information will be included for publication.

The researchers are applying for a Certificate of Confidentiality from the National Institute of Health (NIH). This Certificate will protect the investigators from being forced to release any research data in which I am identified, even under court order or subpoena, without my written consent. This protection does not affect the investigators' legal responsibility to report information about suspected or known sexual or physical abuse of a child or about your

expression of a clear and present danger of harming yourself or others to proper authorities. The Certificate does not prevent me or a member of your family from voluntarily releasing information about myself or my involvement in this study.

Financial Information: I will receive $20 cash for session I and $10 for session II. I will receive $10 for each day I complete my phone app, for a total of $70. The total compensation for this project will not exceed $100. If I prefer, I can receive credit for psychology courses, at the rate of $10 per credit.

Withdrawal: Participation in this study is voluntary. I may withdraw from this study at any time without penalty or loss of any benefit to which I would otherwise be entitled to.

Signatures:
The study has been discussed with me and all my questions have been answered. I may direct additional questions regarding study specifics to the investigators. If I have questions about subjects' rights or other concerns, I can contact Dennis Landin, Ph.D., Chairman, LSU Institutional Review Board, (225)578-8692. I agree to participate in the study described above and acknowledge the researchers' obligation to provide me with a copy of this consent form if signed by me.

_____ _____

Participant Signature Date

_____ _____

Signature of Research Assistant Date

_____ _____

Signature of Principal Investigator Date

Appendix 4

Consent form MinTest - Patients

Forespørsel om deltakelse i forskningsprosjektet

Utvikling av et automatisk 'støttesystem' for monitorering av psykose

Bakgrunn og hensikt
Dette er et spørsmål til deg om å delta i en forskningsstudie for å utvikle et automatisk 'støttesystem' for monitorering av psykose. Ved Psykisk helse- og rusklinikken, Universitetssykehuset i Nord-Norge (UNN), pågår det nå et forskningsprosjekt hvor en ser på språkforandringer hos personer med psykoselidelse, for eksempel ved bipolar lidelse og schizofreni. Denne forskningen skal bidra til utviklingen av bedre metoder for oppfølgning av mennesker med slike tilstander. Som en del av dette ønsker vi å la pasienter, behandlere og andre interesserte prøve ut en mobilapp som heter MinTest, og komme med tilbakemeldinger på om den er brukervennlig og om den oppleves som nyttig. I tillegg vil vi undersøke svarene som gis i appen for å se hvordan de varierer fra dag til dag.

Prosjektleder er professor Brita Elvevåg som er tilknyttet UNN, UiT – Norges arktiske universitet og Nasjonalt senter for e-helseforskning. Prosjektets medarbeidere er klinisk psykolog Connie Malen Moen som er stipendiat ved UNN, og lege Terje Holmlund som er stipendiat ved UiT. Ved spørsmål om prosjektet kan Brita Elvevåg kontaktes på tlf. 45783795.

Hva innebærer studien?
Studien innebærer innsamling av taleprøver både fra personer som har en kjent psykoselidelse, og fra andre som ikke har det. Slik kan man danne et referansemateriale og kalibrere programvaren for det norske språket.

For å gjennomføre prosjektet ønsker vi å bruke et spesiallaget program (en mobilapplikasjon, eller «app») hvor du skal løse ulike oppgaver og snakke om forhåndsbestemte tema. Vi kommer til å bruke en enhet med et mobilt operativsystem (for eksempel iPhone eller iPad) til å samle språkdata, som vil bli analysert av et annet dataprogram. Applikasjonen vil stille spørsmål som er relevante for å kartlegge psykiske tilstander. Oppgavene kan ta opptil 30 minutter å gjennomføre, hver dag, i fem dager. Du kan ta pauser eller avbryte underveis dersom du ønsker det.

Mulige fordeler og ulemper
Det er mulig at oppgavene kan oppleves som fremmedgjørende.

Hva skjer med prøvene og informasjonen om deg?
Informasjonen som registreres skal kun brukes slik som beskrevet i hensikten med studien. Alle opplysningene og resultater vil bli behandlet uten navn og fødselsnummer eller andre direkte gjenkjennende opplysninger. En kode knytter deg til dine opplysninger og prøver gjennom en navneliste. Det er kun autorisert personell knyttet til prosjektet som har adgang til navnelisten og som

kan finne tilbake til deg. Etter prosjektslutt skal datamaterialet anonymiseres. Grunnlagsdata vil være lagret på en forsvarlig måte i 10 år etter at forskningsprosjektet er avsluttet (dvs. til 31.10.2027).

De opplysningene som blir brukt i forskningsprosjektet vil bli behandlet konfidensielt og forskerne har taushetsplikt. Prosjektet er godkjent av personvernombudet ved UNN, samt Regional komité for medisinsk og helsefaglig forskningsetikk. Det vil ikke være mulig å gjenkjenne opplysninger om deg i forskningsrapporten som lages på bakgrunn av studien.

Frivillig deltakelse
Det er frivillig å delta i studien og om du ikke delta vil dette ikke få konsekvenser for deg.
Du trenger ikke å bestemme deg om du vil delta i undersøkelsen med det samme – du må gjerne ta en til to dagers betenkningstid før du bestemmer deg. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke til å delta i studien. Du kan bestemme at innsamlede opplysninger ikke skal benyttes i forskningsprosjektet, uten at dette vil få noen konsekvenser for deg. Dersom du ønsker å delta, undertegner du samtykkeerklæringen på siste side. Om du nå sier ja til å delta, kan du senere trekke tilbake ditt samtykke uten å oppgi årsak. Dersom du senere ønsker å trekke deg eller har spørsmål til studien, kan du kontakte Brita Elvevåg på tlf. 45783795.

Ytterligere informasjon om studien finnes i kapittel *A* – *utdypende forklaring av hva studien innebærer.*
Ytterligere informasjon om biobank, personvern og forsikring finnes i kapittel B – *Personvern, biobank, økonomi og forsikring.*

Samtykkeerklæring følger etter kapittel B.

Kapittel A- utdypende forklaring av hva studien innebærer

For å levere inn forskningsdata må du først laste ned en app, medarbeiderne i prosjektet har mer informasjon om hvordan dette kan gjøres. Hvis du ikke har mobiltelefon eller annen enhet fra Apple som bruker et iOS-operativsystem, kan du låne dette fra forskningsgruppen.

I appen vil du løse ulike oppgaver. En type oppgave består i at du får spørsmål som skal besvares muntlig. Du kan for eksempel få spørsmål om hvordan du har det, eller bli bedt om å gjenfortelle en historie du har hørt. Svarene dine lagres i lydfiler, som senere analyseres ut fra innhold og stemmekarakteristikk. Disse lydfilene vil i tillegg bli brukt for å utvikle et system for talegjenkjenning, og dette systemet vil senere bli brukt til automatisert analyse av språk.
En annen type oppgave består i å trykke på skjermen, for eksempel i en spesiell rytme eller ved å tegne et mønster. Disse oppgavene skal teste det man kaller kognitive funksjoner, slik som hukommelse, oppmerksomhet og språk.

Det er viktig for oss å understreke at svarene du gir oss på oppgavene ikke vil gjøre oss i stand til å konkludere noe om din psykiske helse. For å kunne gjøre noe slikt kreves det mer omfattende undersøkelse utført av helsepersonell. Prosjektet har likevel som mål at slike oppgaver skal kunne bidra med viktig informasjon til helsepersonell som følger opp pasienter med psykiske lidelser.

- **Eventuell kompensasjon til deltakere:** Alle deltakere er med i en trekning hvor man kan vinne en Apple iPad Mini2

Kapittel B - Personvern, økonomi og forsikring

Personvern
Opplysninger som registreres om deg er navn, telefonnummer, mailadresse og postadresse. Disse trenger vi for å kunne ha kontakt når du er med som deltaker, og vil slettes når det ikke lengre er behov for opplysningene (senest 31.10.17). I tillegg vil vi at du registrerer noen bakgrunnsopplysninger, slike som alder og kjønn.

Av kontrollhensyn blir grunnlagsdata oppbevart forsvarlig fram til 31.10.2027. Deretter vil data bli slettet. Det er Brita Elvevåg som er ansvarlig for datamaterialet i denne perioden. Instanser som kan tenkes å kontrollere grunnlagsmaterialet er f.eks. forskningsansvarlige, Uredelighetsutvalget for forskning og Helsetilsynet. Formålet er å kontrollere at studieopplysningene stemmer overens med tilsvarende opplysninger i din journal. Alle som får innsyn har taushetsplikt.

Utlevering av materiale og opplysninger til andre
Hvis du sier ja til å delta i studien, gir du ditt samtykke til at det er kun autorisert personell knyttet til prosjektet som kan få avidentifiserte opplysninger utlevert.

Rett til innsyn og sletting av opplysninger om deg og sletting av prøver
Hvis du sier ja til å delta i studien, har du rett til å få innsyn i hvilke opplysninger som er registrert om deg. Du har videre rett til å få korrigert eventuelle feil i de opplysningene vi har registrert. Dersom du trekker deg fra studien, kan du kreve å få slettet innsamlede prøver og opplysninger, med mindre opplysningene allerede er inngått i analyser eller brukt i vitenskapelige publikasjoner.

Økonomi og Norges Forskningsråds rolle
Studien er finansiert gjennom forskningsmidler fra Norges Forskningsråd. Det er ingen interessekonflikter. Studien er godkjent av Regional komité for medisinsk og helsefaglig forskningsetikk.

Forsikring
Institusjonen er selvassurandør, og deltakere dekkes av produktansvarsloven.

Informasjon om utfallet av studien
Du vil få en kopi av denne deltakerinformasjonen og samtykkeerklæringen. Du kan få en kopi av resultatene av forskningsprosjektet når disse er klare. Du må i så fall angi dette på skjemaet under og oppgi en adresse som resultatene skal sendes til.

Samtykke til deltakelse i studien

Jeg har fått skriftlig og muntlig informasjon og er villig til å delta i studien.

(Signert av prosjektdeltaker, sted, dato)

Jeg ønsker å få tilsendt resultatene av forskningsprosjektet (rapporten) når disse er klare (sett ring rundt):

Ja Nei

Adresse som rapporten skal sendes til:

Appendix 4

Consent form MinTest - Healthy volunteers

Forespørsel om deltakelse i forskningsprosjektet

Utvikling av et automatisk 'støttesystem' for monitorering av psykose

Bakgrunn og hensikt
Dette er et spørsmål til deg om å delta i en forskningsstudie for å utvikle et automatisk 'støttesystem' for monitorering av psykose. Ved Psykisk helse- og rusklinikken, Universitetssykehuset i Nord-Norge (UNN), pågår det nå et forskningsprosjekt hvor en ser på språkforandringer hos personer med psykoselidelse, for eksempel ved bipolar lidelse og schizofreni. Denne forskningen skal bidra til utviklingen av bedre metoder for oppfølgning av mennesker med slike tilstander. Prosjektleder er professor Brita Elvevåg som er tilknyttet UNN, UiT – Norges arktiske universitet og Nasjonalt senter for e-helseforskning. Prosjektets medarbeidere er klinisk psykolog Connie Malen Moen som er stipendiat ved UNN, og lege Terje Holmlund som er stipendiat ved UiT. Ved spørsmål om prosjektet kan Brita Elvevåg kontaktes på tlf. 45783795.

Hva innebærer studien?
Studien innebærer innsamling av taleprøver fra personer som ikke har en kjent psykoselidelse, slik at man kan danne et referansemateriale og kalibrere programvaren for det norske språket.

For å gjennomføre prosjektet ønsker vi å bruke et spesiallaget program (en mobilapplikasjon, eller «app») hvor du skal løse ulike oppgaver og snakke om forhåndsbestemte tema. Vi kommer til å bruke en enhet med et mobilt operativsystem (for eksempel iPhone eller iPad) til å samle språkdata, som vil bli analysert av et annet dataprogram. Applikasjonen vil stille spørsmål som er relevante for å kartlegge psykiske tilstander. Oppgavene kan ta opptil 30 minutter å gjennomføre, hver dag, i fem dager. Du kan ta pauser eller avbryte underveis dersom du ønsker det.

Mulige fordeler og ulemper
Det er mulig at oppgavene kan oppleves som fremmedgjørende.

Hva skjer med prøvene og informasjonen om deg?
Informasjonen som registreres skal kun brukes slik som beskrevet i hensikten med studien. Alle opplysningene og resultater vil bli behandlet uten navn og fødselsnummer eller andre direkte gjenkjennende opplysninger. En kode knytter deg til dine opplysninger og prøver gjennom en navneliste. Det er kun autorisert personell knyttet til prosjektet som har adgang til navnelisten og som kan finne tilbake til deg. Etter prosjektslutt skal datamaterialet anonymiseres. Grunnlagsdata vil være lagret på en forsvarlig måte i 10 år etter at forskningsprosjektet er avsluttet (dvs. til når prosjektet avsluttes den 30.06.2027).

De opplysningene som blir brukt i forskningsprosjektet vil bli behandlet konfidensielt og forskerne har taushetsplikt. Prosjektet er godkjent av personvernombudet for forskning (Norsk samfunnsvitenskapelig datatjeneste, NSD). Det vil ikke være mulig å gjenkjenne opplysninger om deg i forskningsrapporten som lages på bakgrunn av studien.

Frivillig deltakelse
Det er frivillig å delta i studien og om du ikke delta vil dette ikke få konsekvenser for deg.
Du trenger ikke å bestemme deg om du vil delta i undersøkelsen med det samme – du må gjerne ta en til to dagers betenkningstid før du bestemmer deg. Du kan når som helst og uten å oppgi noen grunn trekke ditt samtykke til å delta i studien. Du kan bestemme at innsamlede opplysninger ikke skal benyttes i forskningsprosjektet, uten at dette vil få noen konsekvenser for deg. Dersom du ønsker å delta, undertegner du samtykkeerklæringen på siste side. Om du nå sier ja til å delta, kan du senere trekke tilbake ditt samtykke uten å oppgi årsak. Dersom du senere ønsker å trekke deg eller har spørsmål til studien, kan du kontakte Brita Elvevåg på tlf. 45783795.

Ytterligere informasjon om studien finnes i kapittel *A* – *utdypende forklaring av hva studien innebærer.*
Ytterligere informasjon om biobank, personvern og forsikring finnes i kapittel B – *Personvern, biobank, økonomi og forsikring.*

Samtykkeerklæring følger etter kapittel B.

Kapittel A- utdypende forklaring av hva studien innebærer

For å levere inn forskningsdata må du først laste ned en app, medarbeiderne i prosjektet har mer informasjon om hvordan dette kan gjøres. Hvis du ikke har mobiltelefon eller annen enhet fra Apple som bruker et iOS-operativsystem, kan du låne dette fra forskningsgruppen.

I appen vil du løse ulike oppgaver. En type oppgave består i at du får spørsmål som skal besvares muntlig. Du kan for eksempel få spørsmål om hvordan du har det, eller bli bedt om å gjenfortelle en historie du har hørt. Svarene dine lagres i lydfiler, som senere analyseres ut fra innhold og stemmekarakteristikk. Disse lydfilene vil i tillegg bli brukt for å utvikle et system for talegjenkjenning, og dette systemet vil senere bli brukt til automatisert analyse av språk.
En annen type oppgave består i å trykke på skjermen, for eksempel i en spesiell rytme eller ved å tegne et mønster. Disse oppgavene skal teste det man kaller kognitive funksjoner, slik som hukommelse, oppmerksomhet og språk.

Det er viktig for oss å understreke at svarene du gir oss på oppgavene ikke vil gjøre oss i stand til å konkludere noe om din psykiske helse. For å kunne gjøre noe slikt kreves det mer omfattende undersøkelse utført av helsepersonell. Prosjektet har likevel som mål at slike oppgaver skal kunne bidra med viktig informasjon til helsepersonell som følger opp pasienter med psykiske lidelser.

- **Eventuell kompensasjon til deltakere:** Alle deltakere er med i en trekning hvor man kan vinne en Apple iPad Mini2

Kapittel B - Personvern, økonomi og forsikring

Personvern
Opplysninger som registreres om deg er navn, alder og kjønn.

Av kontrollhensyn blir grunnlagsdata oppbevart forsvarlig fram til 30.06.2027. Deretter vil data bli slettet. Det er Brita Elvevåg som er ansvarlig for datamaterialet i denne perioden. Instanser som kan tenkes å kontrollere grunnlagsmaterialet er f.eks. forskningsansvarlige, Uredelighetsutvalget for forskning og Helsetilsynet. Formålet er å kontrollere at studieopplysningene stemmer overens med tilsvarende opplysninger i din journal. Alle som får innsyn har taushetsplikt.

Utlevering av materiale og opplysninger til andre
Hvis du sier ja til å delta i studien, gir du ditt samtykke til at det er kun autorisert personell knyttet til prosjektet som kan få avidentifiserte opplysninger utlevert.

Rett til innsyn og sletting av opplysninger om deg og sletting av prøver
Hvis du sier ja til å delta i studien, har du rett til å få innsyn i hvilke opplysninger som er registrert om deg. Du har videre rett til å få korrigert eventuelle feil i de opplysningene vi har registrert. Dersom du trekker deg fra studien, kan du kreve å få slettet innsamlede prøver og opplysninger, med mindre opplysningene allerede er inngått i analyser eller brukt i vitenskapelige publikasjoner.

Økonomi og Norges Forskningsråds rolle

Studien er finansiert gjennom forskningsmidler fra Norges Forskningsråd. Det er ingen interessekonflikter. Studien er godkjent av Regional komité for medisinsk og helsefaglig forskningsetikk.

Forsikring
Institusjonen er selvassurandør, og deltakere dekkes av produktansvarsloven.

Informasjon om utfallet av studien
Du vil få en kopi av denne deltakerinformasjonen og samtykkeerklæringen. Du kan få en kopi av resultatene av forskningsprosjektet når disse er klare. Du må i så fall angi dette på skjemaet under og oppgi en adresse som resultatene skal sendes til.

[Et automatisk 'støttesystem' for monitorering av psykose – 3. mai, 2016]

Samtykke til deltakelse i studien

Jeg har fått skriftlig og muntlig informasjon og er villig til å delta i studien.

(Signert av prosjektdeltaker, sted, dato)

Jeg ønsker å få tilsendt resultatene av forskningsprosjektet (rapporten) når disse er klare (sett ring rundt):

Ja Nei

Adresse som rapporten skal sendes til: