Pervasive Computing Technologies for Ambulatory Cognitive Assessment

Hafiz, Pegah

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Hafiz, P. (2020). *Pervasive Computing Technologies for Ambulatory Cognitive Assessment*. DTU Health Technology.

Ph.D. Thesis
Doctor of Philosophy

DTU Health Tech
Department of Health Technology

Pervasive Computing Technologies for Ambulatory Cognitive Assessment

Pegah Hafiz

Kongens Lyngby 2020

Department of Health Technology
Digital Health Section
Technical University of Denmark

Ørsteds Plads, Building 345C
2800 Kongens Lyngby, Denmark
healthtech-info@dtu.dk
www.healthtech.dtu.dk

Summary

Affective disorders cause mood disturbance in individuals and the main types are Bipolar Disorder (BD) and Major Depressive Disorder (MDD). Cognitive impairments in patients with affective disorders may indicate remission of symptoms, which can adversely affect quality of their daily tasks. Neuropsychological tests are the standard tools that are administered to the patients when they come for a follow-up visit. Such tests assess various cognitive domains including memory, attention, processing speed, and executive functions. The follow-up visits are scheduled on a need basis and are administered in a silent room at a clinic by a trained staff. However, clinics have faced lack of resources as the number of patients with affective disorders is increasing. Moreover, human cognitive functioning changes from time to time within a given day. Taken together, frequent assessments are essential to monitor patients' cognitive functioning over time for timely diagnosis and early treatments.

Two pervasive computing technologies were designed, implemented, and evaluated to 1) deliver a patient-administered cognitive assessment tool and 2) assess cognitive functioning in individuals' free-living conditions. Internet-based Cognitive Assessment Tool (ICAT) is a Web-based tool that automatically calculates cognitive test scores regarding verbal memory, working memory, and psychomotor speed. In particular, ICAT utilises speech recognition technology in verbal memory tasks. Hence, patients can take the tests at home without receiving any assistance from a clinician. Ubiquitous Cognitive Assessment Tool (UbiCAT) is a wearable computing technology to 1) assess individuals' attention, working memory, and executive functions over time using three smartwatch-based apps and 2) collect multivariate sensor data including activity and sleep features for digital phenotyping.

Three studies were conducted with ICAT to evaluate usability, feasibility, and concurrent validity of the test scores. The findings of these studies demonstrated high usability and significant validity of the test scores when compared with gold-standard neuropsychological tests. Three studies were also conducted with UbiCAT to investigate usability, validity, and feasibility of this tool in individuals' free-living conditions. As such, high usability of the UbiCAT and a strong correlation coefficient between UbiCAT and standard computerised cognitive tests were obtained. In addition, concurrent validity of the UbiCAT cognitive test scores was demonstrated when compared with neuropsychological tests in a population of healthy controls and patients with BD. Our findings also proved feasibility of UbiCAT in accordance to the participants' Global Positioning System (GPS) data, which showed that cognitive test performance measures were statistically the same in indoor and outdoor places. Supervised learning methods were applied on a dataset including one-week daily observations of cognitive, behavioural, and physiological features of controls and patients with BD for digital phenotyping. As such, Extreme Gradient Boosting (XGBoost) model gave the highest performance and a set of digital phenotypes were derived from this model, showing that individuals' time in bed, daily step counts, number of daily missed counts in the cognitive test sessions, and average of daily executive functioning are the most important features in determining their mental health diagnosis.

Overall, ICAT and UbiCAT are two pervasive computing technologies designed and implemented to provide ambulatory cognitive assessments. The findings of the studies conducted in this thesis demonstrated usability and feasibility of these tools as well as significant validity of their cognitive test scores.

Sammenfatning

Affektive lidelser, såsom depression og bipolar lidelse, er kendetegnet ved at påvirke patientens stemningsleje. Kognitiv funktionsnedsættelse for patienter med affektive lidelser kan indikere forværring i lidelsen og kan negativt påvirke deres dagligdag. Neuropsykologisk test af patienter er ofte en standard procedure når de kommer i ambulant efterbehandling. Sådanne test vurderer forskellige kognitive områder, inklusive hukommelse, opmærksomhed, arbejdshukommelse, og eksekutive funktioner. Disse test skal planlægges, kræver en trænet klinikker til at udføre dem, og et afskærmet og dedikeret lokale i klinikken. Men sundhedsvæsnet står med begrænsede ressourcer samtidig med, at antallet af patienter med affektive lidelser vokser. Endvidere er der et behov for at vurdere patientens kognitive evner også uden for klinikken og i dagligdagen. Af disse grunde er det således relevant at udvikle nye metoder og teknologier, som nemt kan teste patientens kognitive funktionsevne på en mere kontinuerlig måde, hvilket igen kan hjælpe til tidlig diagnose og behandling.

I denne afhandling præsenteres to teknologier til neuropsykologisk vurdering af patienter med affektive lidelser, som dels fokuserer på at patienten selv kan administrere en test og dels, at test kan finde sted i patientens dagligdag uden for klinikken. Det første værktøj – som kaldes Internet-based Cognitive Assessment Tool (ICAT) – er et web-baseret værktøj med fem korte kognitive test. Dette værktøj udnytter moderne stemmegenkendelsesteknologi til vurdering af verbal hukommelse og kan automatisk udregne et testresultat. På denne måde kan patienter tage en test selv, for eksempel derhjemme uden assistance fra en klinikker. Det andet værktøj – kalder Ubiquitous Cognitive Assessment Tool (UbiCAT) – er en bærbar (Eng: “wearable”) teknologi som dels kan indsamle kognitiv funktionsvurderinger over tid ved brug af tre små kognitionstest og dels, indsamler sensor data, herunder data om fysisk aktivitet og søvn.

ICAT har været genstand for tre studier som har vurderet teknologiens brugervenlighed, brugbarhed, samt validitet. Disse studier viste, at teknologien var meget brugervenlig og havde signifikant validitet sammenlignet med eksisterende ‘state-of-art’ metoder. UbiCAT var ligeledes genstand for tre studier som vurderede teknologiens brugervenlighed, brugbarhed, validitet, samt værktøjets anvendelighed under anvendelse i brugerens dagligdag. Teknologien blev vurderet til at være meget brugervenlig og der kunne påvises en stærk korrelation mellem test resultaterne fra UbiCAT og standardiserede kognitive tests. Ydermere blev validiteten af UbiCAT fastlagt både i en gruppe af raske forsøgspersoner samt i en gruppe af patienter med bipolar lidelse. Ved at studere forsøgspersonernes lokationsdata (GPS) kunne det vises, at testresultaterne var valide både når de blev indsamlet indendørs såvel som udendørs.

Maskinlæringsalgoritmer (Eng: 'Supervised Learning') blev brugt til at analysere data indsamlet fra en gruppe af raske forsøgspersoner og fra en gruppe af patienter med bipolar lidelse. Denne analyse viste, at Extreme Gradient Boosting (XGBoost) metoden var den bedste til at identificere hvilke parametre, som gav den bedste forudsigelse af kognitive evne. Disse parametre viste sig at være hvor længe en person befandt sig i sin seng, antal daglige skridt, hvor mange gange forsøgspersonen undlod af udføre sine kognitive test, og den gennemsnitslige test score.

Sammenlagt konkluderer afhandlingen, at ICAT og UbiCAT er to værktøjer som er velegnet til at kunne vurdere personers kognitive evner uden for klinikken og i deres dagligdag. Resultaterne fra studierne viser, at disse to værktøjer er meget brugervenlige, er brugbare, og har en høj validitet sammenlignet med standard metoder.

Preface

This Ph.D. thesis was accomplished at Digital Health Section, Department of Health Technology, Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree.

The Ph.D. project was a part of Technology-Enabled Mental Health (TEAM) programme funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 722561. The Ph.D. project was supervised by Jakob E. Bardram, Professor, Ph.D., DTU Health Tech and Lars V. Kessing, Professor, M.D., Copenhagen University.

The thesis is prepared as a report of the Ph.D. project, which resulted in 6 scientific articles. The work was conducted at DTU Compute (June 2017-January 2019), DTU Health Tech (February 2019-May 2020), and secondments at University College Dublin, Ireland (August-November 2018) and Psychiatric Center Copenhagen, Region Hospital (January-March 2020).

Kongens Lyngby, May 31, 2020

Pegah Hafiz

Acknowledgements

This thesis could not be accomplished without the supports I received from my mentors, family, and friends. First of all, I would like to express my gratitude to my main supervisor Prof. Jakob Bardram for supporting my ideas, giving me freedom in developing my Ph.D. project, numerous practical feedback to our studies, several helps in academic writing, and most importantly giving me a unique opportunity to grow as a researcher. Many thanks to my co-supervisor Prof. Lars Kessing who provided an enjoyable opportunity for me to conduct my clinical studies in Psychiatric Center Copenhagen, Region Hospital. Special thanks to Prof. Kamilla Miskowiak for giving me invaluable insights into the design of digital cognitive tools and our clinical studies. I would like to thank Dr. David Coyle and Erin Finnerty who hosted me at University College Dublin during my external research stay, and several Ph.D. fellows at O'Brien Center for Data Science. I'm also grateful to Prof. John Paulin Hansen for numerous practical inputs to my studies regarding cognitive psychology.

During my secondment at Psychiatric Center Copenhagen, I earned invaluable experience in conducting interdisciplinary research together with psychologists and psychiatrists. I wish to thank a number of medical doctors and research assistants at Psychiatric Center Copenhagen who helped me in recruiting my study participants: Andreas Elleby Jespersen, Kia Obenhausen, Sharleny Stanelous, Ida Seeberg, Marc Nielsen, Helena Bøgh, Viktoria Damgaard, Isabella Røen, Emilie Poulsen, Nana Eskestad, Nynne Müller, Daniel Kjærgaard, and Sarah Møller.

Thanks to Alban Maxhuni for co-authorship, inputs to my studies, and brainstorming sessions with his generous popcorn.

Steven Jeuris for several practical suggestions, finding participants for my studies, discussing the challenges of being an HCI researcher, and coffee times with a brief assessment of the coffee beans. I hope you fix DTU allocation website some day.

Jayden Khakurel for creating a pleasant atmosphere in the office. I will definitely miss our laughters and the times we used to go for sushi.

Likewise, thanks to my Ph.D. fellows at CACHET namely Raju Maharjan and Devender Kumar for our enjoyable collaboration on the Humoris project, Tanya Bafna for our common interest regarding cognitive science and its awesome challenges, and my nice colleagues, Janos Richard Pekk, Jonas Busk, and Kevin Doherty. I would like to appreciate the assistance of several students including Bjorgvin Hjartson and Mads Hesseldahl for promoting research projects at CACHET, Malte Kampmark for graphical designs, and Katarzyna Żukowska for assistance in the ICAT project.

Words cannot describe how much I'm grateful to the emotional support from my parents, Mahdi and Elizabeth, and my sister, Mahboobeh. Without your love, care, and positive energy, I couldn't handle the ups and downs of doing a Ph.D.

Many thanks to Leyla and Sholeh for treating me like their own daughter in Copenhagen. Likewise, thanks to several of my Iranian friends around the world who constantly encouraged me.

I would like to thank my roommates for several workout sessions, BBQ events, and house parties. You guys are incredibly awesome and loud especially during the lockdown period.

As a Marie-Curie fellow, I benefited greatly by the financial support of my Ph.D. project awarded by Europe Union's Horizon 2020, Technology Enabled Mental Health-ITN program. I would like to thank them for giving a fantastic opportunity to me in dissemination of my research at several international conferences and publication of my scientific articles. Last but not least, I wish to thank my study participants, without whom the contributions of my thesis wouldn't be possible.

Contents

List of Figures

List of Tables

Nomenclature

2-CRT Two-Choice Reaction Time

ADHD Attention Deficit Hyperactivity Disorder

ANOVA Analysis of Variance

API Application Programming Interface

ASR Automatic Speech Recognition

AUC Area under the Receiver Operating Characteristic Curve

BACS Brief Assessment of Cognition Schizophrenia

BD Bipolar Disorder

CAMCI Computer Assessment of Memory and Cognitive Impairment

CANS-MCI Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment

CANTAB Cambridge Neuropsychological Test Automated Battery

CARP CACHET Research Platform

CNSVS Central Nervous System Vital Signs

CST Color-Shape Test

CR Consonant Repetition

DART Danish Adult Reading Test

DLL Delayed List Learning

DSST Digit-Symbol Substitution Test

EMA Ecological Momentary Assessment

ESM Experience Sampling Method

GPS Global Positioning System

HAMD Hamilton Depression Rating Scale

HC Healthy Control

HCI Human-Computer Interaction

HRV Heart Rate Variability

ICAT Internet-based Cognitive Assessment Tool

ImPACT Immediate post-concussion assessment and cognitive testing

JS Java Script

KNN K-Nearest Neighbour

LL List Learning

MARS Mobile Application Rating Scale

MCI Mild Cognitive Impairment

MDD Major Depressive Disorder

MyCQ MyCognition Quotient

NASA-TLX NASA Task Load Index

NPV Negative Predictive Value

PC Personal Computer

PCA Principal Component Analysis

PDA Personal Digital Assistant

PDQ-D-5 Perceived Deficits Questionnaire— Depression, 5-item

PPV Positive Predictive Value

PSSUQ Post Study System Usability Questionnaire

PVT Psychomotor Vigilance Test

RBANS Repeatable Battery for the Assessment of Neuropsychological Status

RF Random Forest

RQ Research Question

RT Response Time

SCIP Screen for Cognitive Impairment in Psychiatry

SD Standard Deviation

SVM Support Vector Machines

THINC-it THINC-Integrated Tool

TMT-B Trail-Making Test-part B

TMT Trail Making Test

UbiCAT Ubiquitous Cognitive Assessment Tool

UCD User-Centered Design

UI User Interface

UX User Experience

VF Verbal Fluency

VLT-D Verbal Learning Test-Delayed

VLT-I Verbal Learning Test-Immediate

VMT Visuomotor Tracking

WAIS Wechsler Adult Intelligence Scale

WAIS-LNS Wechsler Adult Intelligence Scale letter-number sequencing

WER Word Error Rate

WMT Working Memory Test

XGBoost Extreme Gradient Boosting

YMRS Young Mania Rating Scale

Part I

Overview

CHAPTER 1

Introduction

1.1 Context and Motivation

Cognitive functioning is one of the core aspects of human mental health and is assessed by examining several cognitive domains including memory, attention, executive functions, and psychomotor speed. Affective disorders disturb individuals' mood [72] and the main types are Major Depressive Disorder (MDD) and Bipolar Disorder (BD). Low mood is a characteristic of patients with MDD. BD, also known as manic-depression, is a chronic illness that causes fluctuations in mood and energy level [62]. It is highly relevant to assess cognitive impairments in patients with BD and MDD as it may indicate the onset of symptoms [8, 38, 87]. In clinical settings, neuropsychological tests are the common practice in psychiatry to assess patients' cognitive functioning. Each assessment session requires a trained staff member to allocate a silent room and a time slot to administer the tests. Moreover, there is evidence that human cognition fluctuates between days and within a single day [77, 100], and follow-up visits are scheduled on a need basis for patients with affective disorders. Thus, patients' cognitive functioning are not assessed frequently, indicating a need for taking the cognitive tests outside clinics.

Recently, mobile cognitive test batteries have been designed and implemented to address this need using extensive resources for cognitive assessments. Such tools make it possible to assess patients' cognition remotely, for instance at home. However, feasibility of such pervasive cognitive assessment tools for frequent and remote assessments of patients with affective disorders has not been adequately explored.

1.2 Background

The first attempt to develop portable devices for cognitive assessment was performed in 1982 by Wilkinson and Houghton who utilized cassette recorders to measure simple reaction times in 10 minutes [102]. Folkard and Monk mentioned that only one cognitive function could be measured with a cassette recorder device and pointed to the time-consuming nature of scoring manually using paper-and-pencil tests [27]. In 1985, Folkard and Monk utilized computer-based test batteries in their studies and emphasized the promising future of portable cognitive assessment tools [29, 63]. Nowadays, psychiatrists administer computerised cognitive test batteries including

Cambridge Neuropsychological Test Automated Battery (CANTAB) [10], THINC-Integrated Tool (THINC-it) [39], CogState [18], and Brief Assessment of Cognition Schizophrenia (BACS) [6] to assess patients' cognitive impairment.

Usability studies aim to improve the design of a system by proposing the tasks to a sample of real users of the system [23]. Evaluating a tool in a laboratory settings is practical to assess its usability since participants are asked to perform certain tasks while the study leader observes and/or records their interaction. However, laboratory settings limit the context for evaluating *Ubicomp tools*, which are ubiquitous computing technologies that require a different context of use compared with laboratory-based tools. *Ambulatory assessments* have been performed using mobile devices such as Personal Digital Assistant (PDA) and smartphones in clinical psychology for various purposes, particularly monitoring [94]. Thus, studies are run in users' *free-living conditions* to evaluate the design of a novel Ubicomp tool [71]. Empirical studies are often conducted in the field of Human-Computer Interaction (HCI) to directly or indirectly observe an evidence. *Feasibility* of a novel tool is often evaluated in an empirical research to investigate viability of the tool. So far, validity of some mobile cognitive assessments tools against the gold-standard neuropsychological tests have been investigated [9, 46, 68, 88]. Behavioural features have been recently analysed for digital phenotyping with smartphones in BD (e.g., [25, 26]) and MDD (e.g., [73, 74]). As such, mobile cognitive assessment tools are capable of collecting behavioural, contextual, and physiological data in conjunction with cognitive test measures to determine digital phenotypes of human mental health.

1.3 Problem Statement

The common practice in psychiatry is to administer neuropsychological test batteries to the patients during clinical visits. Neuropsychological tests often include paper-based and (or) computerised tests. Despite their acceptable accuracy in identifying cognitive impairments, the current procedure in administrating neuropsychological tests imposes three key problems. First, neuropsychological test administration is resource demanding. Due to the increasing number of patients with mental disorder, the number of patients exceeds available resources including trained clinical staff and silent rooms to administer the tests in the clinics. Furthermore, the clinician who administers a test is responsible for monitoring and instructing the participant during the test session. Hence, resource allocation is currently burdensome for psychiatrists and psychologists due to the lack of patient-administered cognitive assessment tools.

Second, neuropsychological tests are often administered in a controlled environment, which is an influential factor in determining validity of the cognitive tests [11, 60]. Previous work suggested taking cognitive tests outside clinics to assess cognitive functioning more frequently and in individuals' free-living conditions [48, 89]. Moreover, research shows that individuals' alertness [77], working memory [28], and executive skills [51] vary within a single day. Therefore, frequent assessments of the cognitive functioning within a day is crucial in achieving reliable test results.

Third, follow-up visits do not provide an opportunity to sense and collect behavioural, contextual, and physiological measures for digital phenotyping. A recent work [16] aimed to determine the digital phenotypes of cognitive functioning during a one-week study and administered neuropsychological tests to the participants as their baseline measure. Passive phone-interaction data of the participants was collected during the study. A limitation reported in this paper was that the neuropsychological tests were administered once per participant, at the beginning of the study while daily frequent testings throughout the study could have provided more insights into their findings. Such a limitation highlights the need for a ubiquitous assessment tool to capture individuals' daily cognitive functioning in conjunction with their mobile sensor data in their free-living conditions.

1.4 Research Questions

The research questions addressed in this thesis are the following:

- **RQ1:** What is the design of a patient-administered cognitive assessment tool for affective disorders?
- **RQ2:** What is the design of an ubiquitous cognitive assessment tool that allow for cognitive assessment in free-living conditions?
- **RQ3:** What is the feasibility of using such cognitive assessment tools for patients with affective disorders in the clinic and in free-living conditions?

1.5 Research Methods

The research method utilised in this dissertation was adapted from the "Triangulation in HCI" methodology by Mackay and Fayard [58]. Figure 1.1 shows the activities performed on three levels: Theory, Design, and Observation. The rest of this section provides an overview of the tasks performed to deliver two pervasive computing technologies for cognitive assessment.

1.5.1 Computerised Cognitive Assessment Tool

SCIP is a paper-and-pencil screening tool to assess cognitive impairment in patients with affective disorders [69]. This tool has five short tests supporting brief assessment of cognition and formed the theoretical basis for the design of the first tool of this thesis called Internet-based Cognitive Assessment Tool (ICAT). A literature review was also performed to identify validated patient-administered tools for mental disorders and to review the design of previous cognitive test batteries. Five major tasks were performed sequentially to design, implement, and evaluate cognitive tests of the ICAT as shown in Figure 1.2. A User-Centered Design (UCD) approach [2]

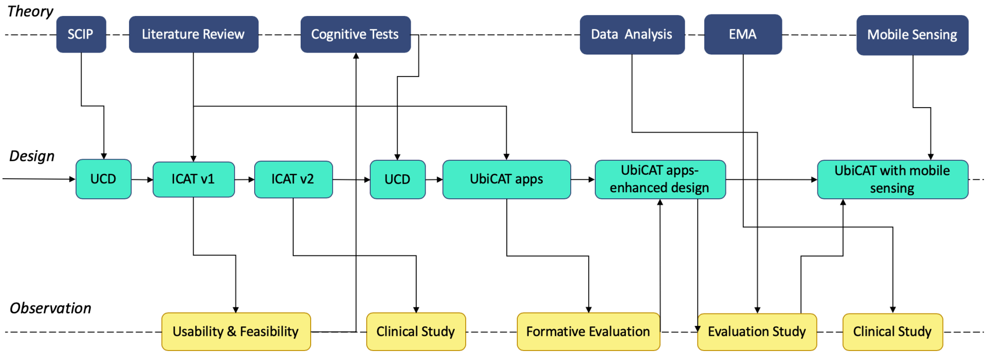


Figure 1.1: The activities performed at each level are presented. (SCIP: Screen for Cognitive Impairment in Psychiatry; EMA: Ecological Momentary Assessment; UCD: User-Centered Design; ICAT: Internet-based Cognitive Assessment Tool; UbiCAT: Ubiquitous Cognitive Assessment Tool).

was adopted to design ICAT in an interdisciplinary team through several meetings. Speech recognition technology was utilised for automatic scoring of the verbal memory tasks in ICAT during remote examinations. ICAT v1 underwent two studies namely usability and feasibility studies to measure participants' psychometric factors and *concurrent validity* of the ICAT test scores against SCIP. Concurrent validity is one of the methods used for criterion validity and it is measured by applying correlation analysis between the test results of a new and an existing tool [55, 64]. Followed by that, we developed ICAT v2 and conducted a clinical validation study with healthy controls and patients with BD.

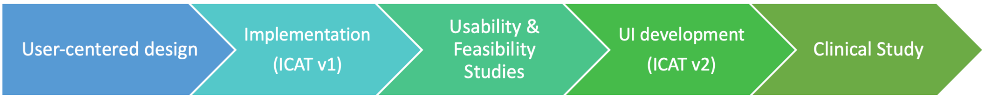


Figure 1.2: Major tasks performed to build and evaluate ICAT.

1.5.2 Smartwatch-based Cognitive Assessment Tool

A literature review of previous mobile cognitive assessment tools was performed to identify suitable neuropsychological tests for implementation on smartwatches. As such, the second tool called Ubiquitous Cognitive Assessment Tool (UbiCAT) was designed together with three domain experts using a UCD approach. Figure 1.3 shows the tasks performed to design, implement, and evaluate UbiCAT as a smartwatch-based tool. Three studies were performed to evaluate UbiCAT cognitive test measures. First, a formative evaluation study was conducted to examine preliminary design of the cognitive tests in UbiCAT, implemented as standalone smartwatch-based apps. Second, an evaluation study was performed to compare the cognitive test performance measures calculated by UbiCAT with their corresponding measures calculated by standard computerised cognitive assessment tools. Last, a clinical feasibility study with controls and patients with BD was conducted to 1) evaluate concurrent validity of the UbiCAT cognitive tests, 2) investigate feasibility of the UbiCAT in a one-week *'in-the-wild'* study, and 3) identify digital phenotypes of human mental health by applying supervised learning methods on a dataset including daily observations of passive mobile sensing and cognitive test results calculated by UbiCAT.

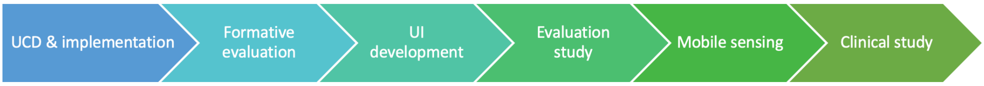


Figure 1.3: Major tasks performed to create and evaluate UbiCAT.

1.6 Contributions

This thesis generally contributes to the design and evaluation of pervasive computing tools for cognitive assessments and more specifically, patient-administered cognitive test batteries and ubiquitous cognitive assessment tools. The main contributions of this thesis are outlined as follows:

1. **Design, implementation, and evaluation of a patient-administered cognitive assessment tool for patients with affective disorders.** A Web-based cognitive assessment tool called ICAT is built and three studies are conducted in two phases with healthy controls and patients with BD to assess perceived usability, feasibility of the tool and speech recognition technology, and concurrent validity of the ICAT test scores.

2. **Design, implementation, and evaluation of a tool for ubiquitous cognitive assessments in users' free-living conditions.** A smartwatch-based cognitive assessment tool is designed and implemented called UbiCAT for continuous assessment of individuals' key cognitive functions outside clinics. Then, three empirical studies are conducted both in a lab and *'in the wild'* to explore subjective human factors and objective cognitive performance measures, feasibility, and concurrent validity of the UbiCAT cognitive tests.

3. **Digital phenotypes of individuals' mental health diagnosis in their free-living conditions.** UbiCAT collects objective measures of human cognitive performance measures. Meanwhile, mobile data including activity and

sleep features are passively collected via wearable and mobile sensors. This research investigates the associations between daily observations of activity and sleep data and cognitive test performance measures. 'State-of-the-art' supervised learning methods are trained and tested on such data to determine the best performing model, from which a set of digital phenotypes of human mental health diagnosis are derived in a population of healthy controls and patients with BD.

1.7 Included Papers

Chapters 9 and 10 present the papers related to the ICAT and UbiCAT, respectively. A brief overview of the papers are presented as follows.

The paper presented below briefly explains the initial User Interface (UI) design of ICAT and was presented at ACM Digital Health Conference 2018 to discuss the opportunities to enhance the UI of this tool:

Article I

Hafiz P, Miskowiak KW, Kessing LV, Bardram JE. **Design and implementation of a web-based application to assess cognitive impairment in affective disorder.** In Proceedings of the 2018 International Conference on Digital Health 2018- Apr 23 (pp. 154-155). (see Section 9.1)

The system design of ICAT and findings of the usability, feasibility, performance of the speech recognition technology, and concurrent validity of the test scores are published in the following paper:

Article II

Hafiz P, Miskowiak KW, Kessing LV, Jespersen AE, Obenhausen K, Gulyas L, Żukowska K, Bardram JE. **The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study**. JMIR formative research. 2019;3(3):e13898. (see Section 9.2)

A formative evaluation study was performed to evaluate preliminary UI design of the smartwatch-based apps in UbiCAT with five participants who had a background in HCI or design and innovation, and to explore the extent to which participants adopt wearables. The findings of this study can be found in the paper presented below:

Article III

Hafiz P, Bardram JE. **Design and formative evaluation of cognitive assessment apps for wearable technologies.** In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers 2019 Sep 9 (pp. 1162-1165). (see Section 10.1)

An empirical evaluation study was conducted to compare the test performance measures calculated by UbiCAT with the corresponding measures obtained from two standard computer-based cognitive assessment tools. The following paper presents the findings of this study and discusses the differences between smartwatch-based and computer-based test performance measures:

Article IV

Hafiz P, Bardram JE. **The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study**. JMIR mHealth & uHealth. (see Section 10.2)

The empirical study introduced above for the Article IV also collected participants' subjective usability rating of the UbiCAT apps, and the cognitive load perceived in the N-back tests. The following paper aims to 1) explore the correlations between perceived human factors and cognitive measures obtained from UbiCAT and 2) group participants on the basis of their demographics using an unsupervised method to discover the similarities between their perceived human factors:

Article V

Hafiz P, Maxhuni A, Bardram JE. **Analysis of Perceived Human Factors and Participants' Demographics during a Cognitive Assessment Study with a Smartwatch.** To appear in: Proceedings of the 8th IEEE International Conference on Healthcare Informatics. (see Section 10.3)

A clinical feasibility study was conducted on UbiCAT with healthy controls and patients with BD. The article below mainly reports 1) feasibility and concurrent validity of the UbiCAT cognitive tests, 2) associations between daily cognitive functioning and wearable sensor data including activity and sleep features, and 3) digital phenotypes of human mental health diagnosis derived from supervised models of bipolar and healthy groups:

Article VI

Hafiz A, Miskowiak KW, Kessing L, Maxhuni A, Bardram JE. **Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls** (see Section 10.4)

1.8 Thesis Overview

Chapter 2 presents and discusses related work on the Web-based and mobile cognitive assessment technologies which have been validated and their feasibility have been previously studied. **Chapter 3** presents the design and implementation of ICAT as well as the key findings from the studies conducted with this tool. The key findings of Chapter 3 are taken from Article I and Article II. **Chapter 4** explains the process in which UbiCAT apps (cognitive tests) was designed and implemented. The main findings of Article III obtained from a formative evaluation study are also included in this chapter since this study was performed to evaluate the design of UbiCAT apps. **Chapter 5** describes two empirical studies conducted to evaluate 1) UbiCAT test performance measures against standard computer-based tests and 2) feasibility of the cognitive test measures of UbiCAT and concurrent validity of the test scores. Some of the key findings of Articles IV and Article VI are presented in this chapter. **Chapter 6** presents the analysis performed on 1) subjective human factors and objective test measures obtained from the first empirical study with UbiCAT and 2) daily cognitive and mobile data to classify healthy and bipolar groups and then, to identify digital phenotypes of human mental health diagnosis. This chapter outlines the findings of Article V and several results of Article VI. **Chapter 7** discusses the main findings of this research regarding the following items:

1. Usability and feasibility of ICAT in calculating cognitive test scores and use of speech recognition technology for automatic scoring of verbal memory tasks, and concurrent validity of this tool against SCIP.

2. Usability of the UbiCAT, feasibility of taking cognitive tests of this tool in individuals' free-living conditions, and concurrent validity of the cognitive tests against neuropsychological tests.

3. Utilising an Ecological Momentary Assessment (EMA) approach for collecting individuals' daily cognitive test performance measures using UbiCAT as well as passive sensing of activity and sleep features in order to extract digital phenotypes of human mental health diagnosis.

The chapter then explains how the outcomes of the studies conducted with ICAT and UbiCAT outperform the findings of previous related work. Followed by that, the chapter provides an outline of some core limitations to the present research as well as some pointers for future work relevant for the HCI and Ubicomp communities. Finally, **Chapter 8** concludes the thesis by explaining how each research question of this thesis was addressed through the main findings, which were obtained from the empirical and clinical studies conducted with ICAT and UbiCAT.

CHAPTER 2

Related Work

A literature review was performed to identify computerised and mobile cognitive assessment tools. Characteristics of the tools are outlined in Section 2.1. Validation and feasibility studies conducted on the computerised and mobile cognitive assessment tools are presented in Section 2.2.

2.1 Overview of Cognitive Assessment Tools

Relevant databases including *PubMed, ACM, and IEEE* were searched to find previous cognitive assessment tools with a peer-reviewed evidence. Note, cognitive training apps with no peer-reviewed evidence were excluded. The tools and applications along with their target users and cognitive domains are presented in Table 2.1. Of the N=19 tools and applications, eleven are Web-based cognitive test batteries *(see #1-11 in Table 2.1)*. Cambridge Neuropsychological Test Automated Battery (CANTAB) is the most well-known cognitive test battery that was built more than 30 years ago. This tool includes validated cognitive tests to examine four key cognitive domain of human cognitive functioning namely attention, processing speed, executive functions, memory and emotional and social cognition. CANTAB can be administered for screening the patients who suffer from cognitive impairments and elderly population with dementia or Alzheimer's disease. 'Mindstreams', Computer Assessment of Memory and Cognitive Impairment (CAMCI), and Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment (CANS-MCI) are particularly developed for assessing cognitive impairment in elderly people. BACS and MyCognition Quotient (MyCQ) assess cognitive functioning of schizophrenic patients and Immediate post-concussion assessment and cognitive testing (ImPACT) is particularly developed for athletes with concussion. 'CogState' is a comprehensive test battery that utilises the card games for assessing and screening various mental disorders including depression, Alzheimer's disease, concussion, schizophrenia, epilepsy, and multiple sclerosis test. Central Nervous System Vital Signs (CNSVS) has seven tests and has been administered to the patients who suffer from BD, MDD, Attention Deficit Hyperactivity Disorder (ADHD), Alzheimer's disease, schizophrenia, substance abuse, and epilepsy. Of the Web-based tools, THINC-it is the only tool that is designed and built specifically for patients with MDD.

Relevant smartphone and wearable applications *(see #12-19 in Table 2.1)* were implemented to provide ambulatory cognitive assessments of various patients including dementia, delirium, MDD, methamphetamine users, and healthy individuals without any mental illness. Color-Shape Test (CST), 'MOBI-COG', and 'iVitality' applications are practical for assessing Mild Cognitive Impairment (MCI) particularly in dementia. 'CognitionKit' has been recently developed by Cambridge Cognition [1]. So far, this tool has shown validated measures for assessing cognitive impairment in patients with MDD but it was mainly built for various target groups. 'Cognition Toolkit' measures *in-the-wild* alertness using three tests and it has been evaluated in empirical studies.

2.2 Validation and Feasibility Studies

CANTAB has been administered to affective disorder patients (for example, [85, 97]), and uses the following tasks to assess cognitive impairments in these patients [14]: *One touch stockings of Cambridge targeting (executive function), Delayed Matching to Sample (visual memory), Emotion recognition task, Spatial Working Memory, and Rapid Visual Information Processing, Cambridge Gambling Task (decision making).* THINC-it has four cognitive tests namely 'Spotter', 'Symbol check', 'Code Breaker', and 'Trails', which were designed to detect cognitive impairments in patients with MDD. This tool also subjectively measures cognition using Perceived Deficits Questionnaire— Depression, 5-item (PDQ-D-5). Total duration of taking cognitive tests of the THINC-it is 15 min. A moderate concurrent validity of the total composite scores of this tool was demonstrated with the patients who had MDD *($r = 0.539$, $p < .001$)* [61]. Another study conducted on THINC-it [39] showed variant convergent validity of the tasks of this tool (between 0.19 and 0.74) in a population of N=100 adults without any mental disorder. Concurrent validity of the cognitive test scores of CNSVS was also examined in a study with patients who had MDD [31]. Their findings showed significant coefficients for processing speed *(p=0.0153)* and attention *(p=0.0013)*. 'CogState' measures were evaluated in [18] that revealed impairments in attention and verbal memory and learning. However, no difference was found between psychomotor speed, visual attention, and working memory of patients with MDD compared with healthy controls.

Table 2.2 provides an evidence for validation of each mobile cognitive assessment tool as well as summarising their studies including participants, cognitive tasks, and key findings. A description of this studies are presented as follows. One of the smartphone apps of a research platform called 'iVitality' implements the following cogntiive tests: *Memory-Word, Trail Making, Stroop, Reaction Time, and N-Back.* Scores calculated by Stroop and Trail Making Test (TMT) tests in 'iVitality' correlated significantly with neuropsychological tests *(r=0.5 and r=0.4, respectively)*, indicating validity of these cognitive tests.

[1] `https://www.cognitionkit.com`

Table 2.1: Overview of the related computerised and mobile cognitive assessment tools.

#	Tool	Reference	Device	Target/Core Aspect	Cognitive Domain
1	CANTAB	Robbins et al. [70]	PC- tablet	Dementia	Attention and processing speed, executive functions, memory, emotional and social cognition
2	CNSVS	Gualtieri et al. [31]	PC	Mental disorders especially MDD	Verbal and visual memory, psychomotor speed, information processing speed
3	Cogstate	Westerman et al. [101]	PC-tablet	Various mental disorders	Psychomotor speed, alertness, visual memory, working memory, verbal memory,executive functions
4	THINC-it	Mcintyre et al. [61]	PC-tablet	MDD	Attention, working memory, and executive function
5	BACS	Atkinds et al. [6]	Tablet	Schizophrenia	Verbal memory, working memory, motor function, processing speed, verbal fluency, executive functions
6	MyCQ	Domen et al. [21]	Tablet	OCD, MDD, schizophrenia	Psychomotor speed, attention, and episodic memory
7	NIH Toolbox	Weintraub et al. [98]	PC	Epidemiological studies	Episodic memory, executive function, attention, processing speed
8	Mindstreams	Dwolatzky et al. [24]	PC	Dementia	Memory, executive functions, visual spatial, verbal function, attention, processing speed, motor skills
9	CAMCI	Saxton et al. [76]	PC	MCI	Attention, verbal memory, visual memory, working memory, executive function
10	CANS-MCI	Tornatore et al. [91]	PC-tablet	MCI especially Alzheimer's disease	Visuospatial ability and spatial fluency, executive control immediate and delayed recall, language fluency
11	ImPACT	Lovell et al. [56]	PC-tablet-smartphone	Concussion	Word memory, sustained attention, reaction time, problem solving
12	CST	Brouillette et al. [9]	Smartphone	Dementia	Processing speed and attention
13	DelApp	Weir et al. [99]	Smartphone	Delirium	Working memory
14	Neurophone	Pal et al. [68]	Smartphone	Methamphetamine users	Working memory, executive functions, inhibitory control
15	CognitionKit	Cormack et al. [15]	Apple Watch	MDD	Working memory
16	Cognition Toolkit	Dingler et al. [20]	Smartphone	Not specified	Attention
17	MOBI-COG	Nirjon et al. [65]	Smartphone	Dementia	Memory, executive function
18	iVitality	Mara et al. [59]	Smartphone	Dementia	Working memory, executive function, processing speed, N-back
19	iHope	iHope Inc. [43]	Smartphone	Not specified	Executive function

CST is mainly practical for elderly people. Validation study showed a moderate correlation between *CST* and global cognition with Mini-Mental State Examination *(r=0.52)*, Digit Span *(r=0.43)*, Trail Making Test *(r=-0.65)*, and Digit Symbol Test *(r=0.51)* but no significant coefficient was obtained for verbal fluency tests. ‘DelApp’ is a smartphone-based app, and it was evaluated against ‘Edinburgh Delirium Test Box’ as a gold-standard computer-based tool. Analysis showed that the test scores calculated by ‘DelApp’ was statistically the same as the gold-standard tool *(p=0.41)*. ‘Neurophone’ is a smartphone-based tool, and it was tested against a computer-based tool with methamphetamine users. ‘Neurophone’ implements N-back and the scores calculated for this test could be compared with standard tools. On the other hand, the scores calculated by Stop Signal test could not be compared with standard tools due to the different scales used for mobile and computer-based tests. Stroop test in ‘Neurophone’ was implemented using speech recognition but it did not have adequate accuracy. As such, this test in ‘Neurophone’ could not be compared with the standard Stroop test.

Table 2.2: Overview of the validation studies with relevant mobile tools.

Application	Participants	Cognitive tasks	Key findings
iVitality [46]	Healthy (N=151)	Memory-Word, TMT, Stroop, Reaction Time, and N-back	Stroop and TMT correlated moderately with the conventional tests
CST [9]	Healthy without dementia (N=57)	Processing speed & attention task	Correlation between CST scores and global cognition with MMSE, digit span, TMT, and digit symbol
DelApp [88]	Delirium patients (N=20)	Not specified	Cognitive test results similar to Edinburgh Delirium Test Box
Neurophone [68]	Healthy (N=20)-methamphetamine users (N=16)	N-back, Stroop, Stop Signal	Stop Signal test results could not be compared; N-back test on both platforms were similar.

Feasibility of mobile cognitive assessment tools were evaluated using an EMA approach including subjective and objective measures. Table 2.3 summarises previous studies that reported feasibility of a mobile tool with cognitive tests. Below, these studies are explained in detail.

PDA was utilised to conduct a study by administrating 2-back tests three times per day for a duration of six days [30]. Cognitive conditions were also evaluated subjectively with a questionnaire including eight items. Participants of this study were N=20 epilepsy patients. Nokia phones were utilised to measure memory and attention twice each day for two weeks in a population including drinkers. In addition, their self report of alcohol consumption were collected [90]. Alcohol consumption was predicted in drinkers using an inhibitory control, that was assessed two times daily via a stop signal task on smartphones [45]. A study compared the scores obtained in laboratory with the scores in uncontrolled conditions using a memory task in N=26 participants [89]. An analysis showed no significant difference between participants' performance in both conditions. Subjective reports of gloomy, fatigue and tension feelings, disturbance level, and locations were also collected. Results showed no significant impact of mood and noise on the test performance measures. An EMA-based study were run using mobile cognitive tests with N=60 elderly people who were not diagnosed with dementia, along with their daily activities [4]. According to their findings, no impact of socializing, general and physical activities were found on their semantic memory in a three-hour time period. Validity and reliability of brief smartphone-based cognitive tests compared with lab-based tests were demonstrated for a period of two weeks per participant [78]. Participants were notified five times per day to take the cognitive tests targeting working memory and perceptual speed.

A study on 'CognitionKit' were conducted to administer N-back tests to N=30 patients with MDD [15]. Participants reported their daily self-reported mood through three short questions on an Apple watch and took an N-back test up to three times per day during six weeks. Their findings showed moderate correlation between daily self-reports and depression questionnaires and daily cognitive assessments with standard tests for patients with MDD. The authors did not report any association between behavioural features (step counts and mood) and cognitive test performance measures, which were collected throughout the study. Alertness of N=12 individuals were examined in an *in the wild* study using a mobile toolkit including three cognitive tasks to measure alertness [20]. The duration of study was on average nine days per participant. The authors also collected time of day, participants' subjective sleep duration and quality, alertness, and preference in terms of cognitive tasks. Their findings revealed effectiveness of Psychomotor Vigilance Test (PVT) and Go No-Go tasks in detecting homeostatic process, and PVT and Multiple Objective Tracking in circadian variations. No significant effect of sleep duration and sleep quality was found on participants' cognitive test performance.

'PsyMate' app was utilised in a study for frequent daily assessment of working memory and processing speed along with self-reports of mood, sleep quality, location, activity, social company and physical status [17]. This study did not show any impact of sleep quality, mood, location, or activity stress on both cognitive test results. A clinical feasibility study with patients who had MDD was conducted using 'iHope' app to measure executive functions daily and to collect self-reports of sleep quality and duration, anxiety, and depression [42]. Participants at the baseline were tested with Hamilton Depression Rating Scale (HAMD) questionnaire to determine severity of depression. HAMD scores correlated with anxiety, depression, and poor sleep quality.

Table 2.3: Overview of the feasibility studies with mobile cognitive assessment tools.

Study	**Instrument**	**Sample**	**Freq.**	**Duration**	**Tests**
Frings et al. [30]	PDA	Epilepsy patients (N=20)	3 daily	6 days	N-back (N=2)
Tiplady et al. [90]	Cellphone	Drinkers (N=38)	2 daily	14 days	Number-pair, matching, memory scanning, sustained attention
Timmers et al. [89]	Smartphone	Healthy (N=26)	4 daily	1 day	Letter span
Sliwinski et al. [78]	Smartphone	Adults (N=219)	5 daily	14 days	Symbol search, dot memory
Abdullah et al. [1]	Smartphone	Students (N=40)	2 daily	40 days	PVT
Dingler et al. [20]	Smartphone	Students (N=12)	1-6 daily	2-13 days	PVT, go no-go, multiple object tracking
Cormack et al. [15]	Smartwatch	MDD (N=30)	3 daily	6 weeks	N-back
Hung et al. [42]	Smartphone	MDD (N=54)	1 weekly	8 weeks	TMT-B, Stroop
Daniels et al. [17]	Smartphone	Healthy (N=49)	8 daily	6 days	Digit symbol substitution, visuospatial working memory

CHAPTER 3

Computerised Cognitive Assessment Tool

ICAT is a patient-administered cognitive assessment tool supporting speech recognition technology for verbal memory tasks. In this chapter, first, the design and implementation of ICAT are presented. Second, main findings obtained from the usability, feasibility and validation studies conducted on ICAT are outlined. The content of this chapter are taken from Article I (see Section 9.1) and Article II (see Section 9.2).

3.1 Design and Implementation

SCIP [69] is a gold-standard tool for patients with affective disorders including five short cognitive tasks namely Verbal Learning Test-Immediate (VLT-I), Consonant Repetition (CR), Verbal Fluency (VF), Verbal Learning Test-Delayed (VLT-D), and Visuomotor Tracking (VMT) tasks. ICAT is the first Web-based cognitive assessment tool adapted from SCIP with minor changes to deliver a patient-administered tool. We formed an interdisciplinary team including computer engineers, User Experience (UX) designers, psychologists, and psychiatrists to design ICAT. Figure 3.1 shows the main tasks performed to design and implement ICAT as explained below.

3.1.1 Team Meetings

We held three brainstorming meetings to discuss the suitable platform on which the tool should be deployed as well as the components required to deliver a patient-administered cognitive test battery. The main components in ICAT are illustrated in Figure 3.2. First, users are provided with a brief overview of ICAT to become more familiar with this tool. Then, an informed consent is displayed before proceeding to the cognitive tasks, providing details of the data types to be collected and how user's data would be handled. Followed by that, general instructions are given to

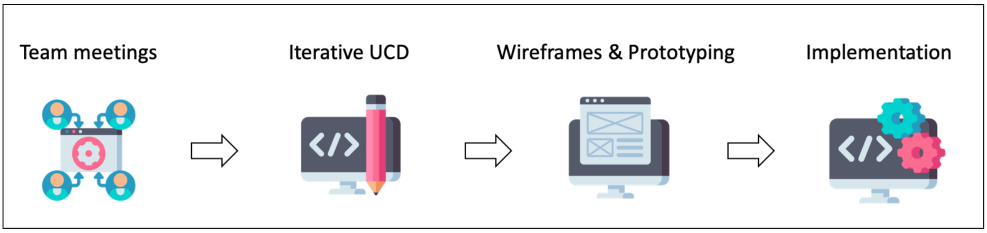


Figure 3.1: Design and implementation procedure of ICAT. The icons in this figure were made by `http://flaticon.com`.

the users. It is essential to familiarize users, in particular patients, with the tasks in ICAT since neuropsychological tests are administered to the patients by a trained staff who is responsible to explain the test instructions while ICAT is designed to be a patient-administered tool. Technical setup including speaker and microphone tests were utilised to ensure that user's device is ready to run the cognitive tests. Next, five cognitive tasks in ICAT are sequentially presented to the users. Finally, user's scores in terms of the number of correct responses achieved in each cognitive test are automatically calculated and displayed to the user.

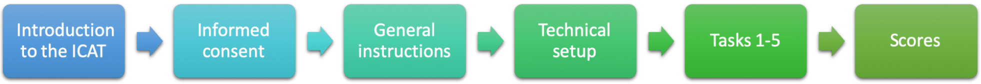


Figure 3.2: General components of ICAT.

3.1.2 Iterative User-Centered Design

Two personas were prepared together with the psychologists and psychiatrists considering the lived experience of the patients with BD. Then, a flowchart was created on the basis of the personas to determine the navigation between components. These items were discussed several times during the team meetings.

3.1.3 Wireframes and Prototyping

UI design was performed by creating wireframes. Low-fidelity mock-ups were created and modified several times during our team meetings. Then, prototypes were created for ICAT and were tested multiple times within the team before implementing the front- and back-end of our tool.

3.1.4 Implementation

The front-end of ICAT was built using React Java Script (JS) framework [3] and the back-end was supported by CACHET Research Platform (CARP). We utilised “Website localization” [75] to implement ICAT in both English and Danish languages. The rest of this section explains the cognitive tasks in Section 3.1.4.1 and the important elements to consider for adapting ICAT from SCIP in Section 3.1.4.2.

3.1.4.1 ICAT tasks

ICAT has five cognitive tasks that are presented sequentially to the users. Characteristics of each cognitive task can be found in Table 3.1. Below, the functionality of each cognitive task is explained:

- **Task 1: List Learning (LL)**– This task is run in three trials. For each trial, a user firstly listens to a list including ten words. Then, the user repeats the words that s/he recalls. The number of correct words are automatically calculated by comparing the word list read to the user with the words recognized by speech recognition technology.

- **Task 2: Consonant Repetition (CR)**– A user listens to a set of letters. Followed by that, the user should complete a short number-sorting task in a limited time duration. Then, the user is required to enter the letters to which s/he listened earlier. These tasks are proposed to the user in several trials with different letters and numbers.

- **Task 3: Wechsler Adult Intelligence Scale letter-number sequencing (WAIS-LNS)**– Similar to task 2, a user first listens to a set of letters and numbers. Then, the user is required to sort the numbers in an ascending order, and the letters in an alphabetical order.

- **Task 4: Delayed List Learning (DLL)**– Similar to task 1, a user is scored on the basis of the number of recalled words captured by speech recognition technology. The word list is the same as task 1 but it only runs in one trial, and the user cannot listen to the word list in this task.

- **Task 5: Visuomotor Tracking (VMT)**– A table including six letters and their matching codes is shown to the users. Thirty random codes are presented and the users should enter the matching codes in thirty seconds.

3.1.4.2 Adapting ICAT from SCIP

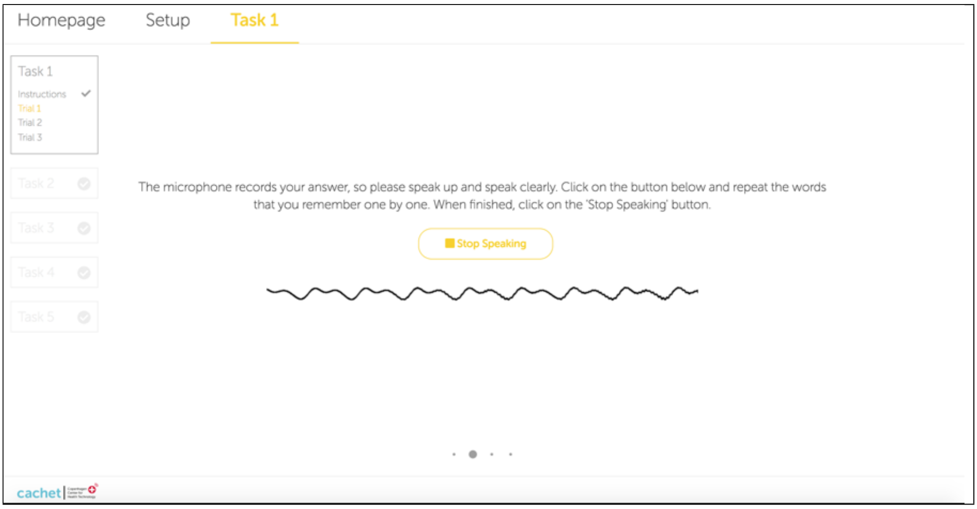
Responses to the cognitive tasks in ICAT were in the form of words, letters, and numbers. The available input methods to enter test responses via a Personal Computer (PC) were a keyboard and a mouse, and speech recognition technology. Five

Table 3.1: Overview of the cognitive tasks in ICAT (Table is copied from [36]).

Features	LL[a]	CR[b]	WAIS-LNS[c]	DLL [d]	VMT[e]
Measure	Immediate verbal recall	Working memory	Working memory	Delayed verbal recall	Psychomotor speed
Criteria	Recalled words for 3 trials	Recalled letters	Sorted sequences	Recalled words	Matched letters
Score	0–30	0–24	0–21	0–10	0–30
Practice set	No	No	Yes	No	Yes

[a] List Learning
[b] Consonant Repetition
[c] Wechsler Adult Intelligence Scale letter-number sequencing
[d] Delayed List Learning
[e] Visuomotor Tracking

issues were considered to computerise SCIP, as a pencil-and-paper tool. First, responses to the verbal memory tasks (LL and DLL) could not be entered manually using a keyboard since 1) such tasks in their conventional setting are administered by a clinician who instructs the patient to repeat the recalled words orally, 2) typing skill of the individuals are not the same as each other, and 3) entering the recalled words involves other aspects of the brain [41] that may distort the recalling procedure. Consequently, we decided to use the Google's Automatic Speech Recognition (ASR) Web service in the verbal memory tasks. Figure 3.3 shows a sample screenshot of the LL task while a user was repeating some words.



The microphone records your answer, so please speak up and speak clearly. Click on the button below and repeat the words that you remember one by one. When finished, click on the 'Stop Speaking' button.

Figure 3.3: A screenshot of the ICAT list learning task (Image is copied from [36]).

Second, the CR task in SCIP presents a number-sorting task to the users immediately after listening to a sequence of letters. The aim of this task is to measure working memory, thus, the mentioned sorting task appears intentionally after the letter sequences to examine user's short-term memory capacity. We used a drag-and-drop component to simulate the sorting task in CR such that users were required to sort some numbers in descending order.

Third, SCIP has a practice set for WAIS-LNS and VMT tasks. Practice sets are essential for the users to learn how to respond during a test. ICAT task 3 and 5 also implement the practice sets of their corresponding task in SCIP. A feedback was given to the users during the practice sets in ICAT to highlight the correct responses in case they made a mistake or to approve their correct responses.

Fourth, the VMT task in SCIP includes a table of letters with their corresponding Morse codes (see Appendix A). Taking this test in SCIP requires the subjects to write down the matching Morse codes of a sequence of thirty letters. Since writing the corresponding Morse codes is faster than typing them (for example, due to the common trouble in finding the dash (-) and dot (.) symbols), we swapped the test stimuli and responses with each other. Moreover, since some people are already familiar with Morse codes, the test difficulty could have been perceived differently by users. As such, users were required to enter the matching letters of a sequence of codes, consisted of circles (○) and asterisk ($*$) symbols. The preliminary UI design of the ICAT VMT task is presented in [35]. We decided to remove a timer from the test to avoid inducing stress to the users (see Figure 1 in Section 9.1).

Last, the VF task in SCIP was replaced with WAIS-LNS for implementation in ICAT since ASR could not be used for the VF task. VF tasks require users to repeat as many words as possible which start with a certain letter. However, ASR converts any random word to the closest word in the dictionary, thus, it was considered unreliable for utilization in ICAT. It should be noted that WAIS-LNS was selected by the psychiatrists as VF and WAIS-LNS measure the same aspect of executive functions.

3.2 Usability and Feasibility Studies

We conducted three studies to evaluate usability and feasibility of ICAT v1. An overview of the usability and feasibility studies on ICAT v1 is depicted in Figure 3.4. First, usability evaluation was performed with $N = 21$ individuals who did not have any mental illness before. Followed by that, another study with $N = 19$ healthy controls were conducted at Psychiatric Center Copenhagen (Region Hospital). The objectives of these studies were to evaluate usability and feasibility of the ICAT v1 with healthy controls and to calculate accuracy of the ASR in the verbal memory tasks (LL and DLL) separately for English and Danish responses.

3.2.1 Participants and Procedure

Study 1 was conducted with individuals who studied or worked at Technical University of Denmark or city of Copenhagen. Participants of study 2 were healthy controls

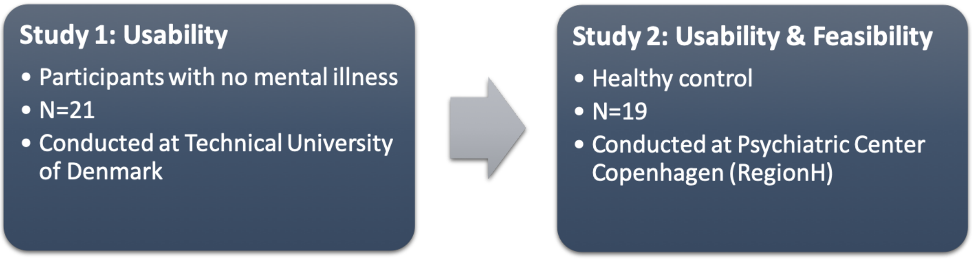

Study 1: Usability
• Participants with no mental illness
• N=21
• Conducted at Technical University of Denmark
Study 2: Usability & Feasibility
• Healthy control
• N=19
• Conducted at Psychiatric Center Copenhagen (RegionH)

Figure 3.4: Overview of the studies conducted on ICAT.

who were recruited from Danish blood bank donors. Procedures of the studies are explained as follows. First, participants demographics were collected. Then, they took the cognitive tests of ICAT with no assistance following the *think-aloud* method [44]. Upon finishing the tests, participants rated their perceived usability with Post Study System Usability Questionnaire (PSSUQ) [52] and took part in a short follow-up interview. The test sessions and the interviews were recorded. Manual transcripts of the participant's recalled words were extracted from their records to calculate accuracy of the ASR during the verbal memory tasks in ICAT. Word Error Rate (WER) measure was utilised by previous work [50, 67, 93], hence, we used this metric to report ASR accuracy using the following formula (equation is copied from [36]):

$$WER = (S + D + I)/N \tag{3.1}$$

N is the total number of the words, D is the number of deletion, S is the number of substitutions, and I is the number of insertions. Upon finishing study 1, we conducted another study at Psychiatric Center Copenhagen with healthy controls to measure usability of ICAT and concurrent validity of the cognitive test scores calculated by this tool against SCIP. Participant took the Danish version of SCIP and ICAT tests in a randomized order. Usability of ICAT was also assessed in this study. To evaluate concurrent validity, Pearson's correlation analysis was applied to the participants' cognitive test scores obtained from ICAT and SCIP tests.

3.2.2 Key Findings

In total, we recruited N=40 individuals. Demographics of the study participants are summarized in Table 3.2. Participants' ratings using the PSSUQ were calculated to obtain their overall score, system usage, information quality, and interface quality. Table 3.3 presents participants' ratings separately for each study, and for Danish and English speaking participants to get better insights from their subjective psychometric factors.

The correlation analysis between ICAT and SCIP test scores obtained in study 2 are reported in Table 3.4. Note, since ICAT tasks implement SCIP version 3, we

Table 3.2: Demographics of ICAT study participants.

Study number	Age	Gender	Native Language	
			English	*Danish*
Study 1	31±12	F=9, M=12	9	12
Study 2	36±15	F=13, M=6	0	19

Table 3.3: Psychometric factors of the PSSUQ to evaluate ICAT usability (Table is copied from [36]).

Factor	Study 1 (N=21)	Study 2 (N=16)	English test (N=25)	Danish test (N=12)
Overall score	4.12±0.46	4.36±0.42	4.25±0.45	4.19±0.45
System usage	4.23±0.53	4.52±0.41	4.39±0.48	4.35±0.45
Information quality	3.86±0.55	4.24±0.58	4.11±0.55	3.84±0.64
Interface quality	4.28±0.62	4.25±0.49	4.16±0.57	4.50±0.45

also used SCIP version 2 with a different word list for the verbal memory tasks to minimize practice. We visualized word accuracy and the number of recalled words for English and Danish words. Figure 3.5 illustrates the performance evaluation for English words. The rest of the results can be found in Article II (see Section 9.2).

3.2.3 Lessons Learned

Usability studies are conducted to investigate where users had trouble in interacting with a system. The records of the test sessions and the semi-structured interviews of the studies on ICAT were analyzed to identify the following issues:

1. Participants generally did not read the instructions of the tests carefully although the instruction sets were organized in a point-by-point style. They often skimmed quickly over the texts and some of them asked questions during the tests which were already addressed in the instructions.

2. The drag-and-drop component in the CR task was not as user-friendly as we expected as some of the participants called it ‘confusing’. In contrast to the counting task in SCIP, this component did not challenge the short-term memory of our participants. Furthermore, participants of the study 2 mentioned that the drag-and-drop component in ICAT was easier than the counting task in SCIP.

Table 3.4: Correlation analysis between ICAT and SCIP test scores obtained from study 2 (Table is copied from [36]).

Cognitive domain	**SCIP task**	**ICAT task**	**R**	**P value**
Verbal learning (SCIP-2^{a}) using ASR transcripts	VLT -I	LL	0.56	**.013**
Verbal learning (SCIP-3^{b}) using ASR transcripts	VLT-I	LL	0.67	**.002**
Verbal learning (SCIP-3^{b}) using manual transcripts	VLT-I	LL	0.66	**.002**
Working memory (SCIP-2^{a})	WMT	CR	-0.12	.63
Working memory (SCIP-2^{a})	WMT	CR	0.11	.65
Executive function (SCIP-3^{b})	VF	WAIS-LNS	0.29	.27
Delayed recall (SCIP-3^{b}) using ASR transcripts	VLT-D	DLL	0.34	.15
Delayed recall (SCIP-3^{b}) using manual transcripts	VLT-D	DLL	0.58	**.009**
Psychomotor speed (SCIP-3^{b})	VMT	VMT	0.71	**.001**
Total score	total	total	0.63	**.009**

[a] SCIP- version 2
[b] SCIP- version 3

3. Some of the participants had difficulty in finding the 'tab' key on their keyboard to navigate between the text boxes in the VMT task in ICAT which might be due to the less frequent use of this key compared to the rest of control keys.

4. Some participants had an unclear and quiet voice. Consequently, some of the words in the verbal memory tasks were not correctly recognized by ASR although participants repeated them.

3.3 Enhancing ICAT

The aforementioned issues presented in Section 3.2.3 led us in enhancing UI of ICAT before conducting a clinical validation study with BD patients and healthy controls. We decided to add audio files of each test instruction set such that users firstly listen to the instructions, then, they read the same instructions in a point-by-point style. The problem with the drag-and-drop component was fixed by utilising ASR in task 2 since we noticed that participants were generally engaged in the speech recognition technology. The issue with finding the 'tab' key in the VMT task was solved by displaying a symbol of this key to the users as shown in Figure 3.6. Therefore, users would proceed only if they find this key on their keyboard. Some of the words were

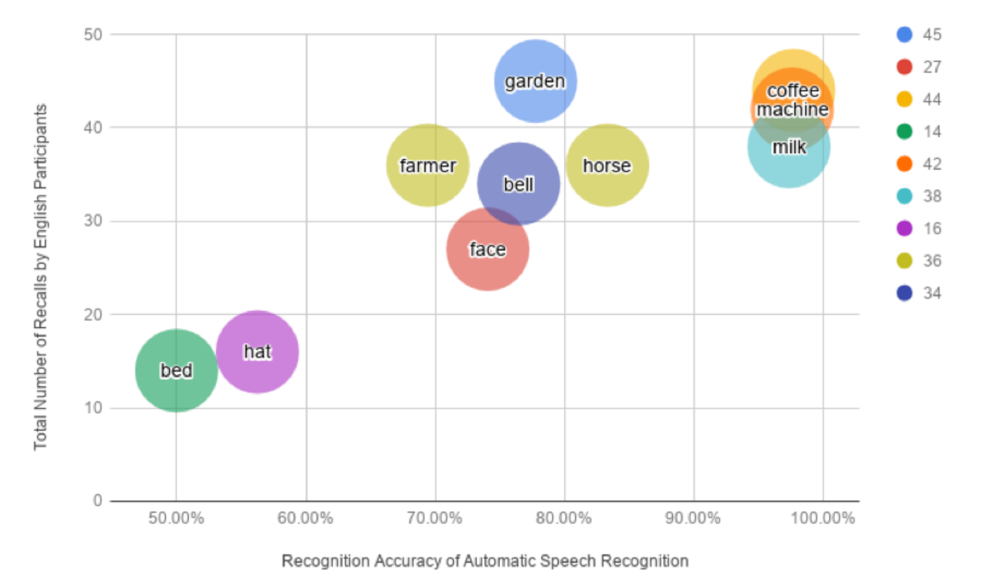
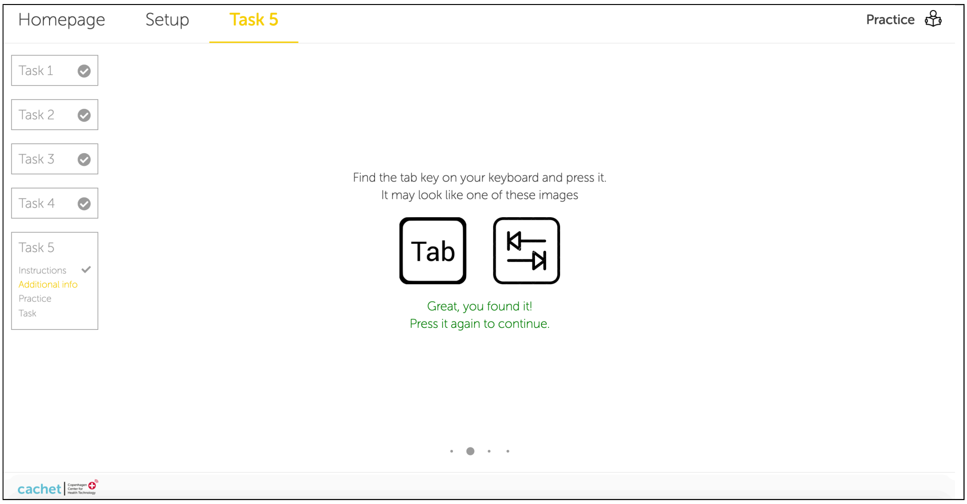


Figure 3.5: Performance of ASR in verbal memory tasks for an English word list (Figure is copied from [36]).

difficult to be recognized by the ASR. We removed such words and selected some of the words from other validated versions of SCIP. ICAT v2 was deployed after applying the mentioned modifications to the UI design. The changes are outlined in Table 3.5.

Table 3.5: Changes applied to the cognitive tasks of ICAT v1.

Task name	**Modification**
List Learning	Replaced few English and Danish words
Consonant Repetition	Automatic speech recognition replaced with drag-and-drop component
Wechsler Adult Intelligence Scale letter-number sequencing	No change
Delayed List Learning	Same as List Learning task
Visuomotor Tracking	Identifying position of the ‘tab’ key prior to the test



Find the tab key on your keyboard and press it.
It may look like one of these images

Great, you found it!
Press it again to continue.

Figure 3.6: Interactive method implemented in ICAT to find the 'tab' key.

3.4 Clinical Validation Study

This section briefly reports preliminary findings of a clinical validation study conducted at Copenhagen Affective Disorder Research Centre, Copenhagen University, Region Hospital. The aim of this study was to evaluate concurrent validity of the test scores calculated by ICAT v2. The finding of this study was presented as a poster at 22$^{\text{nd}}$ Annual International Society for Bipolar Disorders Conference, and a manuscript to report the findings will be prepared later.

3.4.1 Participants and Procedure

Adults with BD who were in full or partial remission and healthy controls were assessed with the Danish version of SCIP and ICAT v2. Participants took ICAT and SCIP in a randomized order. The criteria for identifying their mental health diagnosis was defined on the basis of scores ≤ 14 on the HAMD 17 items and Young Mania Rating Scale (YMRS).

3.4.2 Findings

The preliminary analyses included patients with BD (N= 23) and healthy controls (N=26). Patients displayed cognitive impairment compared with healthy controls as measured by ICAT total score *(t-score = 2.15, p=0.03)*. The analyses revealed a strong association between SCIP and ICAT total scores *(r = 0.73, p < .001)*.

Moreover, the individual cognitive tasks were positively correlated with each other *(r= 0.50-0.73, p < .001)*.

3.5 Chapter Summary

ICAT is a Web-based cognitive assessment tool with short tests adapted from a paper-and-pencil tool called SCIP. ICAT automatically calculates test scores regarding immediate and delayed verbal memory, working memory, and psychomotor speed. In this chapter, UCD process of the ICAT was described which would help future researchers in creating patient-administered tools. Modified elements for implementing cognitive tests in ICAT were presented, indicating the challenges of computerising paper-based tests. The studies conducted with healthy individuals and patients with BD demonstrated acceptable usability, feasibility of the design and implementation of ICAT, and concurrent validity of the test scores calculated by ICAT. Therefore, patients with affective disorders can take the cognitive tests of ICAT at home upon receiving a request from their psychiatrists.

CHAPTER 4

Designing Cognitive Tests for Smartwatches

The idea of creating a smartwatch-based tool for cognitive assessment was mainly raised from the gaps in the literature. First, there is no ubiquitous tool to assess key cognitive functions including attention, working memory, and executive functions. Second, smartphone-based cognitive assessment tools cannot collect physiological data such as sleep. Ubiquitous Cognitive Assessment Tool (UbiCAT) is the first smartwatch-based tool that measures key cognitive functions for frequent assessments during EMA-based studies. This chapter presents the design and implementation of UbiCAT in Sections 4.1 to 4.3, cognitive tests of UbiCAT in Section 4.4, and the key findings of a formative evaluation study in Section 4.5 that verified the preliminary UI design of the UbiCAT apps. The content of Section 4.5 is taken from Article III (see Section 10.1).

4.1 Device Selection

Ubiquitous devices including smartphones and wearables can prompt their users frequently and in various contexts. Compared with the smartphones, wearables impose less limitations with respect to their mobility. For instance, observations show that individuals are not willing to bring their smartphones when they go for a run or walk. Hence, wearables may be more practical to be administered in the EMA-based studies for moment-by-moment assessments. Of the wearables, smartwatches are often worn for visibility and usefulness purposes [13]. The requirements for selecting a smartwatch platform for UbiCAT implementation then become the following;

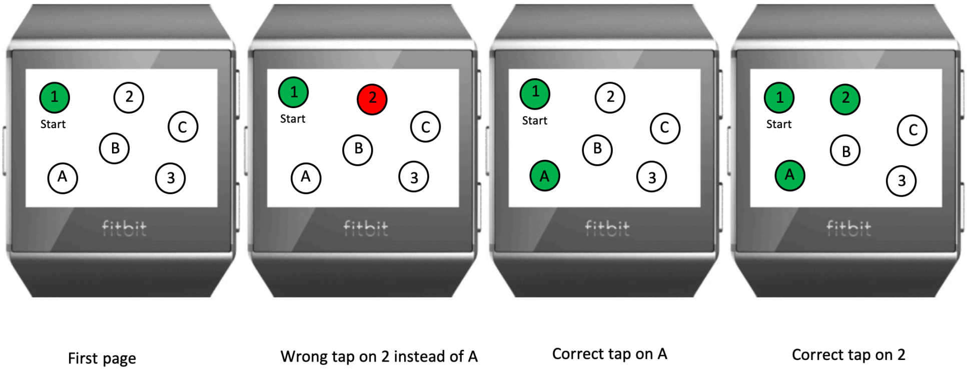
1. the smartwatch should support stand-alone application development, and
2. the Application Programming Interface (API) of the smartwatch should support behavioural, contextual, and physiological data collection.

A review of available smartwatch technologies was performed during October-December 2018. Based on this review, Fitbit (Ionic and Versa) and Apple Watch were the only smartwatches that met the aforementioned requirements. Battery discharge was a crucial factor since in the EMA-based studies frequent recharging is an additional burden imposed to the participants. Since an Apple Watch is often discharged every other day, Fitbit smartwatches was chosen over the Apple Watch. After comparing Fitbit Ionic and Versa with each other, Ionic was selected due to its 1) GPS support and 2) longer-lasting battery life (approximately five consecutive days).

Fitbit API allows researchers to program their own app using JS in 'Fitbit Studio' and publish their app publicly or to limited users, who are given a link to download it. In addition, Fitbit apps can be executed through 'Fitbit Studio' and data can be stored locally on the internal memory of Fitbit smartwatches. Several meetings were held together with three domain experts to discuss essential factors to consider when designing the cognitive tests for smartwatches, and feasibility of implementing the tests.

4.2 Neuropsychological Tests for Implementation

After choosing Fitbit Ionic smartwatches, standard neuropsychological tests were investigated for implementation on Fitbit smartwatches. An impairment in attention, memory, and executive function can bring about problems at work or school [57]. Furthermore, daily fluctuations were previously identified in alertness [77], working memory [28], and executive function [51]. Verbal memory tasks could not be implemented on smartwatches since these tests require a microphone to automatically recognize the recalled words. Besides, the use of microphone in the public areas may not have adequate acceptability as discussed during the meetings with domain experts. A review was performed on the available neuropsychological tests from which mobile tests had been adapted. As such, five tests were initially selected to measure the aforementioned key cognitive functions: N-back [49] to measure working memory, Two-Choice Reaction Time (2-CRT) [22] for attention, and Stroop color-word test [84], Trail-Making Test-part B (TMT-B) [83], and Digit-Symbol Substitution Test (DSST) [53] for executive functions. Of these tests, TMT-B and DSST could not be implemented on Fitbit smartwatches. Figure 4.1 and Figure 4.2 illustrate the mock-ups created for TMT-B and DSST tests, respectively. As can be seen, the circles are too small to capture users' taps on the watch screen. In addition, standard TMT-B test includes 9 pairs of digits and letters. It was not possible to implement all elements of the TMT-B test in a single screen. It should be noted that reducing the number of digit-letter pairs (similar to Figure 4.1) could have led to invalid test results due to insufficient number of test stimuli. Likewise, symbols (shapes) and digits in the DSST could not fit properly, and the size of circles was not large enough to capture user's taps on the screen. Consequently, TMT-B and DSST were removed from the available options such that 2-CRT, N-back, and Stroop tests were implemented on Fitbit Ionic smartwatches.



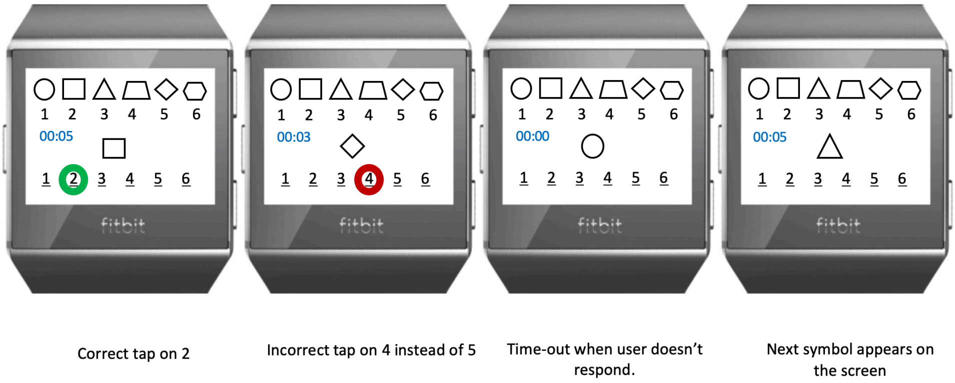
First page

Wrong tap on 2 instead of A

Correct tap on A

Correct tap on 2

Figure 4.1: Mockup created for TMT-B on the Fitbit Ionic.



Correct tap on 2

Incorrect tap on 4 instead of 5

Time-out when user doesn't respond.

Next symbol appears on the screen

Figure 4.2: Mockup created for DSST on the Fitbit Ionic.

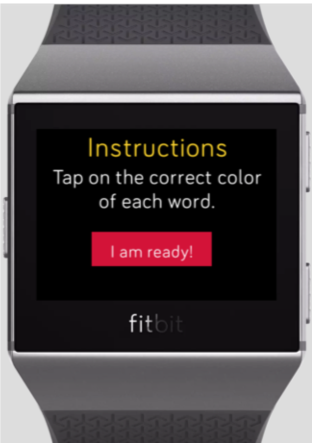
4.3 User-Centered Design and Implementation

Three cognitive tests were selected for implementation on the Fitbit Ionic smartwatch. In this section, UCD process of UbiCAT tests are presented. It should be noted that each cognitive test is a standalone, smartwatch-based app. Thus, the terms 'UbiCAT apps' and UbiCAT tests' are used interchangeably from this point. Design of the elements used in UbiCAT apps is inline with Fitbit design guidelines [1]. Functional prototypes were tested frequently via 'Fitbit Simulator'. Each app was run on a Fitbit Ionic device and tested with several users with various finger sizes to 1) refine navigation between the app views [2] and 2) resize the app buttons to a proper scale to

[1] `https://dev.fitbit.com/guides/design-guidelines`

[2] Each page displayed on a smartwatch is called a 'view'.

capture users' taps. The common components of each UbiCAT app are an instruction set and the actual test. Initially, a single-view instruction was created for each app to summarize the instructions. Figure 4.3 shows an early design of the instructions in Stroop test. However, users found such instructions insufficient, thus, a 'panorama view' was utilised to present more details for each test instruction. The final design of Stroop test instructions using 'panorama view' is shown in Figure 4.4.



Instructions

Tap on the correct color of each word.

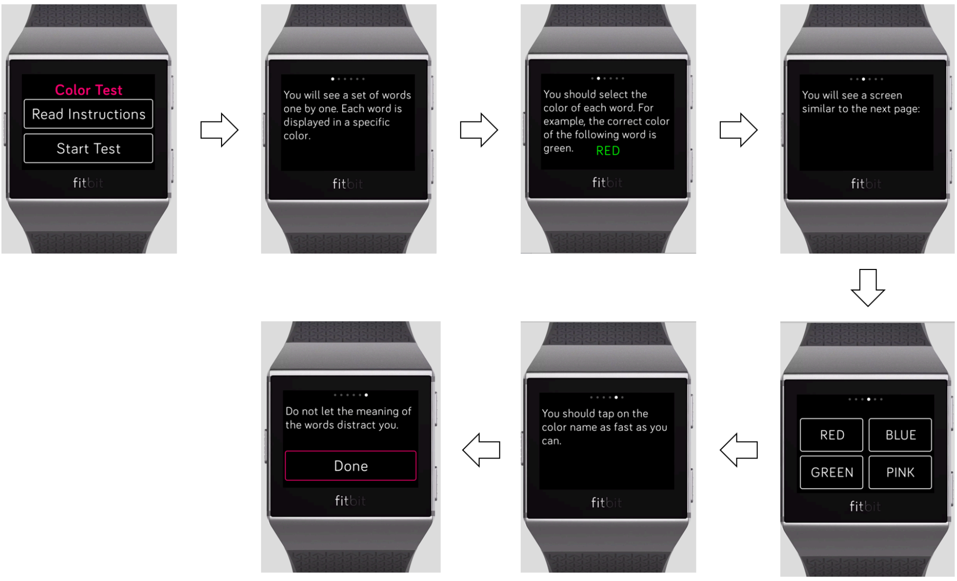
Figure 4.3: Initial instruction design of the Stroop test in UbiCAT.

4.4 UbiCAT Cognitive Tests

This section elaborates on the cognitive tests used for each UbiCAT app along with some screenshots of the apps to clarify their functionality. Each app implements a standard neuropsychological test.

4.4.1 Two-Choice Reaction Time Test

This test measures user's attention and processing speed through user's number of correct responses and Response Times (RTs) to the test stimuli. In UbiCAT 2-CRT app, each view presents a right-hand or left-hand arrow on either left or right side of the screen. Two app buttons appear on both sides of the screen so users are required to tap on the correct direction of each arrow as fast as possible. The time limit to select the direction of an arrow is 2500 ms. Figure 4.5 shows an initial and the final UI design of UbiCAT 2-CRT test.



Color Test

Read Instructions

Start Test

You will see a set of words one by one. Each word is displayed in a specific color.

You should select the color of each word. For example, the correct color of the following word is green. RED

You will see a screen similar to the next page:

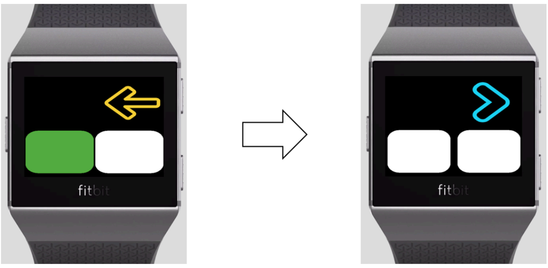
RED BLUE GREEN PINK

You should tap on the color name as fast as you can.

Do not let the meaning of the words distract you.

Done

Figure 4.4: Final instruction design of Stroop test in UbiCAT.



(a) (b)

Figure 4.5: Initial and final design of the 2-CRT test. Part b is copied from [32].

4.4.2 N-back Test

N-back test measures working memory by presenting a sequence of letters. A user should memorize N letter back in the sequence and indicate whether the current letter appeared N letter before or not. Three difficulty levels were considered for the N-back test in UbiCAT such that 1-back, 2-back, and 3-back tasks could run on the smartwatches. Figure 4.6 shows UI development of UbiCAT N-back test. The color of letters was changed from yellow to blue as a user suggested.

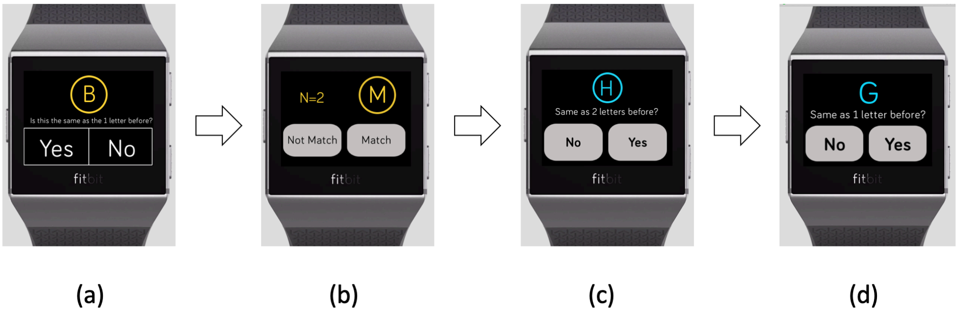


Figure 4.6: N-back user interface development in UbiCAT.

4.4.3 Stroop Color-Word Test

Stroop color-word test measures executive functions. Similar to the classic Stroop test, a set of congruent and incongruent stimuli is displayed to the users in each trial. A congruent stimuli is defined as a color name displayed with the same color as its meaning (for example 'RED') while an incongruent stimuli is a color name displayed in a different color (for example 'GREEN'). The task of users is to select the ink color of each stimuli in a limited time. Figure 4.7 illustrates how the UI design of the Stroop test in UbiCAT improved.

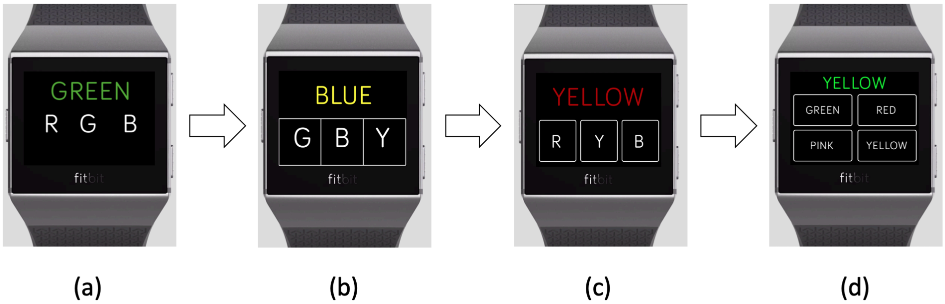


Figure 4.7: Stroop's user interface design improvements.

4.5 Formative Evaluation Study

A study was conducted on an early design of UbiCAT to 1) evaluate the preliminary design of the apps and 2) investigate adoption of wearables by study participants. Three names were assigned to each app in order to simplify memorizing the apps by the participants:

- 2-CRT test → Arrow Test
- N-back test → Letter Test
- Stroop test → Color Test

4.5.1 Participants and Procedure

Five individuals from Technical University of Denmark participated with a background in design and innovation or HCI. This study was performed in five steps per participant as illustrated in Figure 4.8. First, an informed consent form was handed to the participants to inform about data collection and handling of their data. Upon signing the form, participants were asked about their previous experience of using a wearable device. If the participant had already used a wearable device, the reason(s) for quitting or adopting the device were inquired. Then, participants were asked to wear a Fitbit Ionic smartwatch and perform the following tasks for each UbiCAT app:

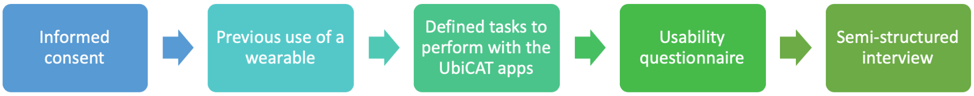

Informed consent
Previous use of a wearable
Defined tasks to perform with the UbiCAT apps
Usability questionnaire
Semi-structured interview

Figure 4.8: Tasks performed during a formative evaluation study with UbiCAT apps.

1. Launch a UbiCAT app on the smartwatch.
2. Read the test instructions in the app.
3. Take a cognitive test with the app.
4. Check your score at the end of the test.

These tasks were recorded and the participant was asked to verbalize their thoughts following the *'think-aloud'* method. Then, participant was asked to fill in a usability questionnaire including seven questions extracted from Mobile Application Rating Scale (MARS) questionnaire [82] concerning three psychometric factors: aesthetics, functionality, and information quality and quantity of the UbiCAT apps. Finally, a semi-structured interview was held with the participant.

4.5.2 Key Findings

In this section, key findings are presented regarding participants' statistics, their perceived usability ratings, and interviews conducted with them. The rest of the findings can be found in Article III (see Section 10.1).

4.5.2.1 Participants' Statistics

The study was run with five participants (1 female, 4 male; age= 28 ± 4.35). One participant held a Ph.D. degree in HCI, three participants were Ph.D. candidates in the field of HCI, and one participant was studying in Design and Innovation program at Master's level.

4.5.2.2 Usability Ratings

Five-point Likert-based scale was used to rate the psychometric factors of the MARS questionnaire. Selected questions of MARS used in this study can be found in Appendix B. Table 4.1 reports participants' ratings for each UbiCAT app.

Table 4.1: Usability ratings of the UbiCAT apps (Table is copied from [32]).

UbiCAT App	**Aesthetics**	**Functionality**	**Information**
Arrow Test	$3.93 \pm .61$	4.6	4.4
Letter Test	$4 \pm .2$	3.3 ± 1.84	2.6
Color Test	$4 \pm .2$	$4.2 \pm .28$	$3.9 \pm .14$

4.5.2.3 Previous Use of Wearable Devices

Three participants did not use any wearable device. Participants were asked about the wearable devices they had used. Two of them used a wearable device for a while. For instance, **P4** used Basic Pick and Apple Watch for several months.

4.5.2.4 Interviews

Participants also talked about an active feedback displayed on the watch screen after responding to a test stimulus. A sample feedback to a correct response in the 'Color Test' given to a user is presented Figure 4.9. Test scores were displayed in the last view of each UbiCAT app. It reported a single number, indicating the total number of correct responses during a test session. According to the meetings held with two domain experts, we decided to show a single number to the users as a low test score could have an adverse effect on the users. However, all study participants were

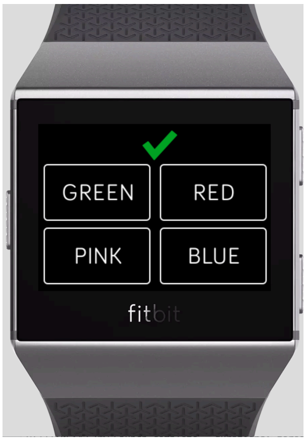


Figure 4.9: Sample feedback to a correct response in UbiCAT Stroop test.

looking for the maximum number of scores in a certain format (for example, 28/30). Therefore, the maximum number was added to the scores.

4.5.3 Post-Study Improvements

The finding of formative evaluation of UbiCAT revealed that the apps are usable and there is no major issue to tackle with. The information quantity and quality of the 'Letter Test' received lower ratings compared to the other apps (see Table 4.1). Therefore, the instructions were enhanced by adding more details. Functionality of the 'Letter Test' also received lower rating which might be due to the inherent difficulty of the N-back test. Upon analyzing the results of this study, minor changes were applied to the UI design of the apps including bigger font sizes and replacing some colors.

4.6 Chapter Summary

Design, implementation, and formative evaluation of the UbiCAT apps were presented in this chapter. Important factors that should be considered by future researchers are outlined as follow. First, device selection is an important part of smartwatch app development for doing a research *'in-the-wild'* since the opportunities of existing smartwatches are not the same as each other. Second, limitations of small screens and brief interactions with smartwatch-based apps are the key factors to consider when designing for smartwatches. UbiCAT apps meet these criteria as it takes less than 2 minutes to take a test with any of the apps, and each event requires a short command. Third, creating paper prototypes for smartwatch apps is less efficient as papers do not provide any opportunity for identifying where app buttons are not sensitive to user's taps on the watch screen. Hence, high-fidelity prototypes are essential for iterative tests with users. Finally, formative evaluation studies explore user's interaction and assist in revising UI design of the apps where necessary before conducting larger studies with more participants.

CHAPTER 5

Empirical Evaluations of UbiCAT

This chapter outlines the findings of two empirical studies conducted on UbiCAT. The first study aimed to compare cognitive performance measures of UbiCAT with standard computer-based cognitive tests and the results are presented in Section 5.1, which are taken from Article IV (see Section 10.2). The second study was conducted in collaboration with Psychiatric Center Copenhagen, Region Hospital. Results regarding feasibility and concurrent validity of the UbiCAT test scores are reported in Section 5.2, which are taken from Article VI (see Section 10.4).

5.1 Comparison with Computer-based Tests

Computer-based cognitive tests implement standard neuropsychological tests to assess cognitive functioning. Smartwatch-based apps have a different interaction space compared with PCs. It is unknown how cognitive performance measures of a test calculated via smartwatches are compared with the measures obtained from a computer-based test. Thus, an empirical evaluation study was performed to compare the cognitive test measures of UbiCAT with computer-based tests.

5.1.1 Metrics and Tools

A search over the existing computer-based tools was performed in the first step considering two criteria: 1) The cognitive tests of such tools should implement standard neuropsychological tests and 2) The tool should allow changes to the test parameters in order to run the cognitive tests with the same parameters on both computer and smartwatch platforms. Consequently, THINC-it [39] and Psytoolkit tests [80, 81] were selected as standard computer-based tools. The THINC-it *Spotter test* implements the choice reaction time test suitable for evaluation against UbiCAT 2-CRT app. Psytoolkit is an open-source toolkit including several standard cognitive tests and allows changing the test parameters. As such, the number of test stimuli and maximum time limit in the N-back and Stroop color-word tests of Psytoolkit were managed to be the same as in UbiCAT.

The usability evaluation of this study was performed with MARS questionnaire (see Appendix B). Perceived cognitive load of the participants during the N-back tasks on both computer and smartwatch was assessed using NASA Task Load Index (NASA-TLX) questionnaire [40]. The sub-scales of NASA-TLX included in this study are mental and temporal demand, overall performance, effort and frustration level. The test sessions with each participant were recorded to 1) check the interactions with UbiCAT cognitive tests (apps) and 2) extract themes from the interviews by performing semantic analysis on their responses.

5.1.2 Participants and Procedure

Healthy individuals (female: 12, male: 9) without previous history of a mental illness were recruited from Technical University of Denmark using snowball sampling method [7]. Participants' education were from bachelor, master, or Ph.D. levels, and they worked or studied in various industries. The tasks performed in this study are depicted in Figure 5.1. As can be seen, the tasks fall into three phases: before, during, and after experiment. A description of each phase is outlined as follow. Each participant was briefed with a short introduction to the study and what s/he should have expected during the test session. Then, an informed consent was handed to the participant. Upon signing the consent form, sociodemographics of the participant was collected including age, gender, higher education level (in years), job title, work or study industry, and the use of dominant or non-dominant hand when wearing a

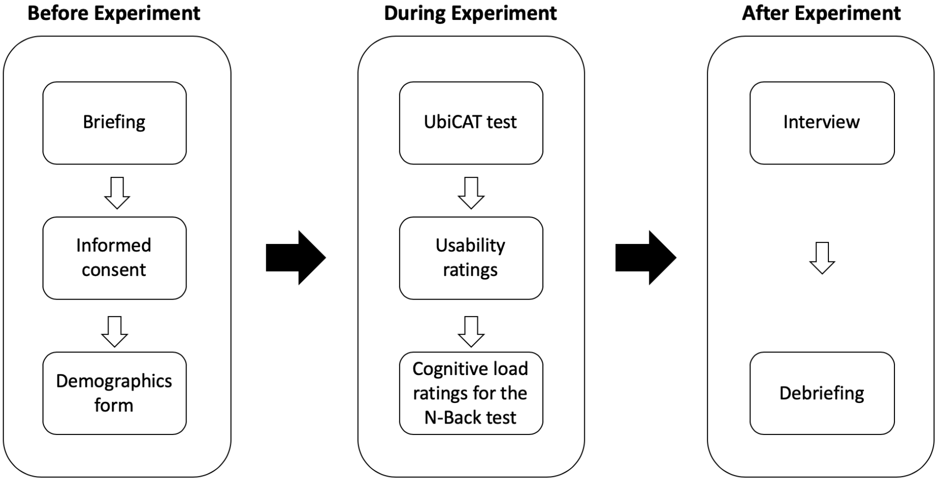


Figure 5.1: Tasks performed for comparing UbiCAT with computer-based tests.

watch. Each participant took a cognitive test both on a computer and Fitbit smartwatch using UbiCAT in a randomized order. Followed by that, the participant was asked to evaluate usability of the UbiCAT app through which s/he took a cognitive test. For the N-back tests on both platforms, participants additionally rated their perceived cognitive load. A short interview was conducted with each participant immediately after phase 2 to investigate where s/he struggled and to improve UI design of the UbiCAT apps. At the end, participants were debriefed about the objectives of this study and further explanations were given to them.

5.1.3 Correlation between Cognitive Tests

Pearson's correlation analysis between total scores (accuracy) of the participants obtained from computerised and UbiCAT tests revealed a significant coefficient *($r= 0.78$, $p<0.001$).* Figure 5.2 shows a positive association as well as the confidence intervals. Next, the association between the performance measures of each test on both platforms were investigated. The average of correct responses obtained from THINC-it 'Spotter test' and UbiCAT 2-CRT correlated significantly with each other. The mean RTs obtained from each N-back task on UbiCAT and Psytoolkit correlated significantly as shown in Figure 5.3. Furthermore, the RTs to Stroop congruent and incongruent stimuli on Psytoolkit and UbiCAT correlated significantly with each other.

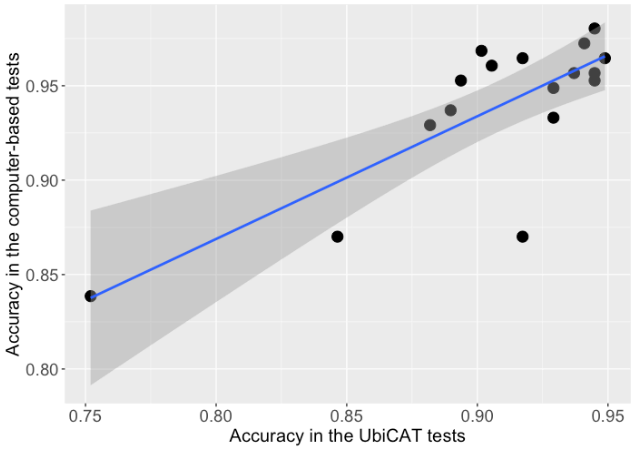


Figure 5.2: Correlation between UbiCAT and computer-based test scores (Figure is copied from [33]).

Analysis of Variance (ANOVA) revealed a significant effect of task difficulty on participants' test performance measures in the N-back tests. The study also demonstrated usability ratings $>$ 4 (out of 5) in terms of aesthetics, functionality, and information quality and quantity. Low discomfort $<$ 3 (out of 7) was reported by

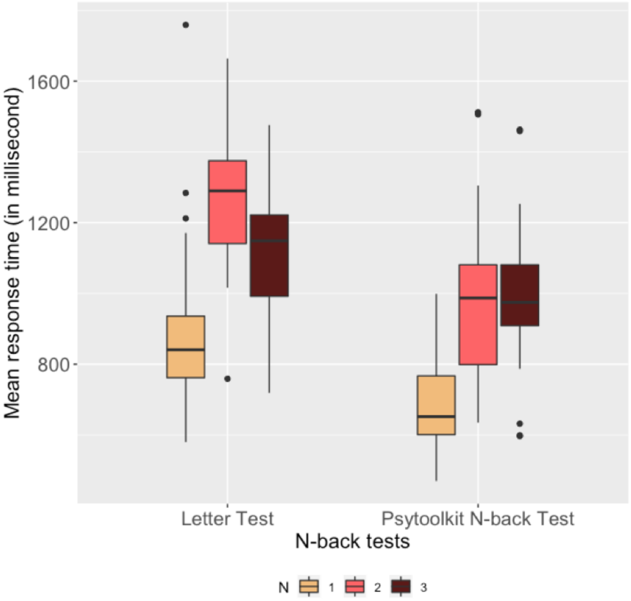


Figure 5.3: Correlation between response times in the N-back test obtained from UbiCAT and Psytoolkit (Figure is copied from [33]).

our participants after taking the tests with the UbiCAT apps. Recorded interviews with the participants were transcribed and seven themes were extracted including *perception about the experiment, input modality, device screen, visual impact, psychological factors, performance, and suggestions for enhancing the UI of UbiCAT.* While majority of the participants preferred taking cognitive tests on smartwatches, a few participants preferred computer-based tests. The rest of the findings can be found in Article IV (see Section 10.2).

5.2 Clinical Validation Study

A controlled clinical study was conducted at Psychiatric Center Copenhagen to evaluate concurrent validity of the UbiCAT test scores against neuropsychological tests. Feasibility of this tool was evaluated by comparing participants' performance measures in indoor and outdoor places using their GPS data, which was collected passively when they took the cognitive tests of UbiCAT. Study procedure and results of the concurrent validity and feasibility of UbiCAT are presented in this section.

5.2.1 Participants and Procedure

Patients with BD and healthy controls were recruited from Psychiatric Center Copenhagen, Region Hospital. Participants came for their follow-up visits and underwent several neuropsychological tests. These tests were selected as the baseline to evaluate concurrent validity of the UbiCAT test scores. Followed by their visits, participants took the cognitive tests of UbiCAT one by one and their results were used to apply correlation analysis between the scores obtained from neuropsychological tests and UbiCAT tests. In total, N=15 participants (6 bipolar patients, 9 healthy controls) were included. Participants' demographics as well as ratings of the YMRS and HAMD are reported in Table 5.1.

Table 5.1: Characteristics of the participants in UbiCAT clinical study (Table is copied from [37]).

Characteristic	**Measure**	**Statistics**	
		Healthy control	**Bipolar patient**
Gender	Female (Nr.)	5	5
	Male (Nr.)	4	1
Age	Mean±SD	34±12	32±6
Years of education	Mean±SD	16±1.8	15±2.04
HAMD	Mean±SD	1.1±1.3	5.2±3.5
YMRS	Mean±SD	0.7±2	2.3±3.2
Verbal Intelligence Quotient	Mean±SD	115±5	108±3

5.2.2 Concurrent Validity of UbiCAT

The neuropsychological tests administered during the follow-up visits included the following tests: TMT parts A and B, Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) coding and digit span [19], WAIS-LNS, and verbal fluency [95]. Scores obtained in these tests were used to extract composite scores for each cognitive domain. Z-transformations of the RBANS coding and digit span and the TMT part A tests were averaged to calculate a score for attention and processing speed. Executive function was obtained by averaging z-transformations of the verbal fluency and TMT part B tests Working memory was calculated directly using the z-transformations of the WAIS-LNS test. Global cognitive scores was calculated to compare with the overall scores of the UbiCAT by averaging z-transformations of the composite scores of attention and processing speed and working memory. Pearson's correlation analysis revealed that participants' scores obtained in UbiCAT correlated significantly with their corresponding neuropsychological tests *(r= 0.58-0.64)*.

Moreover, participants' global cognition calculated using their total scores correlated strongly with each other *(r= 0.77, p<0.001)*. Correlation analysis results are presented in Table 5.2.

Table 5.2: Correlation analysis between neuropsychological tests and UbiCAT test scores (Table is copied from [37]).

Cognitive Function	**Neuropsychological Test**	**UbiCAT Test**	**r**	**p**
Executive functions	Verbal fluency and TMT-B	Stroop's score to incongruent stimuli	0.58	**0.024**
Working memory	WAIS-LNS	N-Back	0.63	**0.011**
Attention and processing speed	TMT-A and RBANS	Choice reaction time	0.64	**0.010**
		Stroop's score to congruent stimuli	-0.11	0.686
Global cognition	Working memory and attention	Composite scores	0.77	<**0.001**

5.2.3 Feasibility of UbiCAT

An EMA approach was used to conduct a one-week study on UbiCAT cognitive tests with healthy controls and patients with BD. Twelve participants took the tests both at indoor and outdoor places according to their GPS data. ANOVA was used to examine the impact of environment on participants' test performance measures. As shown in Table 5.3, participants' measures were statistically the same in indoor and outdoor places. Therefore, UbiCAT is a feasible tool for assessing *in-the-wild* key cognitive functioning of the individuals.

Table 5.3: Impact of indoor and outdoor places on UbiCAT test measures using analysis of variance (Table is copied from [37]).

UbiCAT Test	**Observations**	**Performance Measure**	**Mean Square**	**F**	**p**
2-CRT	249	Median RTs	4505.44	0.45	0.501
		Nr. of correct responses	0.41	0.35	0.553
N-Back	217	Mean RTs	15478.79	0.34	0.560
		Nr. of correct responses	9.38	0.09	0.759
Stroop	227	Mean RTss	14294.25	0.30	0.583
		Nr. of correct responses	2.50	0.57	0.452

Of the participants, N=7 took part in a short interview and mentioned the issues regarding the experiment with UbiCAT tests as well as the frequent places and contexts in which they took the tests. In general, none of the participants reported a problematic issue with the smartwatch and the experiment itself. One participant felt uncomfortable to take a test when people were around. Participants mentioned that they took the tests while *sitting, walking, and standing.* Some participants mentioned that they took the tests *on the bus* as well.

5.3 Chapter Summary

This chapter presented the findings of two empirical studies that would help researchers in designing evaluation studies on their novel pervasive computing tools. Significant correlations were revealed between the scores obtained in UbiCAT and computer-based tests. Similarly, the significant concurrent validity of UbiCAT test scores was demonstrated against neuropsychological tests. Hence, UbiCAT tests calculate valid measures of individuals' cognitive functioning. Feasibility of UbiCAT in a one-week study using an EMA approach was also shown as there was no significant impact of indoor and outdoor environments on the cognitive test performance measures of UbiCAT. Consequently, UbiCAT has valid cognitive test measures and acceptable feasibility for *'in-the-wild'* administration.

CHAPTER 6

Analysis of UbiCAT Cognitive Measures

UbiCAT calculates various cognitive performance measures per cognitive test including number of correct responses, longest scoring streak, RTs, and missed stimuli, that is defined as the number of stimuli to which a user does not respond during a test session. These key performance measures provided opportunities to derive new insights from users' data. The analysis performed to identify the associations between subjective human factors and objective measures of UbiCAT collected from N=21 healthy adults are presented in Section 6.1. Furthermore, an unsupervised learning method was utilised to visualize participants' human factors on the basis of their sociodemographics. This section reports the key findings of Article V (see Section 10.3). The findings of a one-week study on UbiCAT with N=15 patients and healthy controls are presented in Section 6.2 regarding 1) hourly-basis alertness, 2) comparing mobile and cognitive data of patients with healthy controls, 3) correlations between sleep and working memory performance, and 4) digital phenotypes of human mental health extracted from feature raking of a trained model of XGBoost. This section presents several results of Article VI (see Section 10.4).

6.1 Subjective and Objective Measures

Usability and cognitive load are two important human factors. Cognitive load is an essential psychometric factor such that an excessive load inducted by a tool can negatively affect learnability of the tool [47, 86, 103]. Usability metrics of the 2-CRT and 1-back apps in UbiCAT and perceived cognitive loads of the participants in the 1-back tasks formed the subjective psychometric factors while participants' cognitive test performance in these tests were the objective measures of cognition calculate by UbiCAT in terms of RTs, number of correct responses, and longest scoring streaks of the participants. Such data were used to 1) analyze the relationship between subjective and objective measures using correlation analysis and 2) explore the similarities between participants' perceived usability and cognitive load in terms of their sociodemographics. The tasks performed for this study are illustrated in Figure 6.1. Table 6.1 shows participants' sociodemographics including gender, education, age, in-

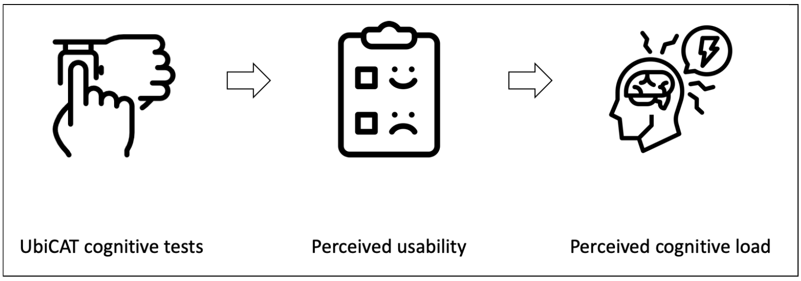


Figure 6.1: Overview of the tasks performed for analyzing subjective and objective measures. The icons in this figure were made by `http://flaticon.com`.

dustry, and job. Perceived usability was calculated by three metrics selected from the MARS questionnaire [82] (see Appendix B). Participant rated their perceived cognitive load during a 1-back task using sub-scales of the NASA-TLX questionnaire. The rest of this section presents the main findings of Article V.

6.1.1 Correlation between Human Factors and Cognitive Measures

Significant correlation coefficients were revealed between participants self-reports of aesthetics and functionality, and the number of correct responses in the 2-CRT test of UbiCAT as shown in Table 6.2. Similar analysis was performed on the 1-back test that showed strong correlation coefficients between participants' perceived functionality of the UbiCAT 1-back app and all of the cognitive performance measures obtained from this test. Table 6.3 presents the correlation coefficients for these analysis.

6.1.2 Clustering based on Sociodemographics

Ward hierarchical clustering method [96] was performed on the participants' usability and cognitive load ratings to cluster them on the basis of their sociodemographics. The cluster analysis revealed a different pattern between the ratings of female and male participants regarding their perceived cognitive load and usability metrics as depicted in Figures 6.2 and 6.3. Key findings of this study are summarised below:

1. Participants' could not accurately assess their own performance during the cognitive tests by comparing their rating a sub-scaled of NASA-TLX and their actual test scores calculated by UbiCAT.

2. Participants from various background rated usability metrics and cognitive load sub-scales differently.

The rest of the figures and results can be found in Section 10.3.

Table 6.1: Sociodemographics of the study participants (Table is copied from [34]).

Variable	**Characteristics**	**Nr. (%)**
Gender	Male	**12** (57.14%)
	Female	**9** (42.86%)
Education	Bachelor degree	**6** (28.57%)
	Master degree	**8** (38.10%)
	Ph.D.	**7** (33.33%)
Age	19-30	**17** (80.95%)
	31-40	**3** (14.29%)
	> 40	**1** (4.76%)
	Mean ± SD	**26.90 ± 5.98**
Industry	Design	**4** (19.05%)
	Research	**4** (19.05%)
	Computer Engineer	**4** (19.05%)
	Construction	**1** (4.76%)
	Education	**1** (4.76%)
	Energy Engineer	**1** (4.76%)
	Food Engineer	**1** (4.76%)
	Healthcare	**3** (14.29%)
	Research	**4** (19.05%)
	Water Engineer	**2** (9.52%)
Job	Student Assistant	**3** (14.29%)
	Bachelor Student	**3** (14.29%)
	Master Student	**5** (23.80%)
	Ph.D Student	**4** (19.05%)
	Postdoctoral Researchers	**3** (14.29%)
	Data Analyst	**1** (4.76%)
	Nurse	**1** (4.76%)
	Project Manager	**1** (4.76%)

6.2 Daily Cognitive Measures and Mobile Data

Unobtrusive active and passive data was collected from the participants in a one-week study with UbiCAT. At the same time, participants took cognitive tests of the UbiCAT three times per day. Sleep data and activity features such as step counts were passively collected via Fitbit API. Such behavioural and physiological data as well as cognitive test measures allowed for collecting daily features to classify patients with BD and healthy controls and extracting important features of human mental health diagnosis that is called digital phenotypes [66]. Table 6.4 gives an overview of the cognitive, behavioural, contextual, physiological data that were collected for this study per participant. Note, GPS data was used to evaluate feasibility of the

Table 6.2: Correlation analysis for 2-CRT test (Table is copied from [34]).

Test	Usability Metrics		
Measure	*Aesthetics*	*Functionally*	*Information*
Mean RT	0.00	-0.11	0.21
Correct responses	0.45*	0.63**	0.38
Longest streak	0.52*	0.63**	0.53*

*p<0.05

**p<0.01

Table 6.3: Correlation analysis for 1-back test (Table is copied from [34]).

Test	Usability Metrics		
Measure	*Aesthetics*	*Functionally*	*Information*
Mean RT	-0.28	-0.66**	-0.43
Correct responses	0.20	0.73***	0.42
Longest streak	0.30	0.75***	0.48*

*p<0.05

**p<0.01

***p<0.001

UbiCAT cognitive tests as presented in Section 5.2.3. The one-week study helped in analyzing various measures of participants' cognitive functioning in relation to their behavioural, contextual, and physiological data. Seven findings are outlined as follows to compare cognitive and mobile features of the controls with the patients:

1. Patients' alertness was lower than the healthy controls according to their median RTs in the 2-CRT test *(t-score= 5.24, p<0.001).*

2. Patients experienced more drops and less rises in their attention compared with the controls according to the ratio of negative and positive alertness (see Figure 6.5).

3. Patients were less capable of responding timely to a test stimulus as their daily missed counts were higher compared with the controls *(t-score=3.24, p<0.001)).*

4. Patients' processing speed was lower compared with the controls according to their lower RTs in the UbiCAT Stroop test *(t-score=1.93, p=0.029).*

5. Patients' sleep duration was significantly higher than the controls *(t-score= 3.68, p<0.001).*

6. Patients stayed more in bed for sleeping compared with the controls *(t-score= 3.46, p=0.001).*

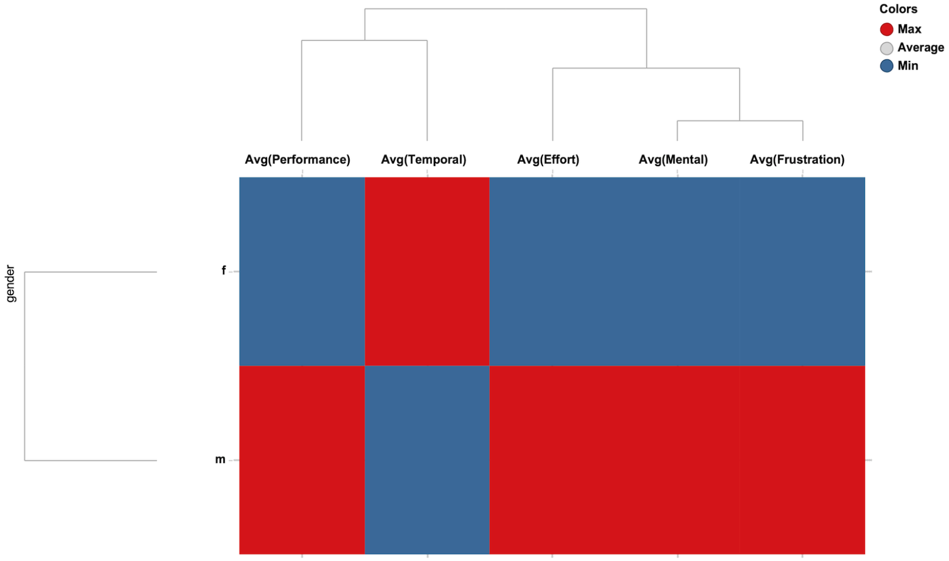


Figure 6.2: Perceived cognitive load of female and male participants in the 1-back task (Figure is copied from [34]).

7. Mobility of the patients was higher than the controls according to their step counts *(t-score= 2.03, p=0.046)*.

6.2.1 Sleep Duration and Working Memory

A dataset was prepared including daily sleep data and the next-day cognitive performance measures, which were averaged through daily test sessions. This dataset had N=74 observations. Correlation analysis was performed on several cognitive test performance measures and sleep duration of the participants. Of the test measures, N-back hit rates as a measure of working memory correlated significantly with sleep duration *(r=0.26, p=0.026)*. It can be inferred that higher sleep duration had a positive association with accuracy of the participants in recognizing correct matches between the letters presented in the N-back tasks.

6.2.2 Digital Phenotypes of Mental Health

Classification of daily observations of the patients with BD and healthy controls was performed by adding activity features to the dataset for sleep analysis. This dataset had N=81 observations including the following features: *time in bed, sleep duration,*

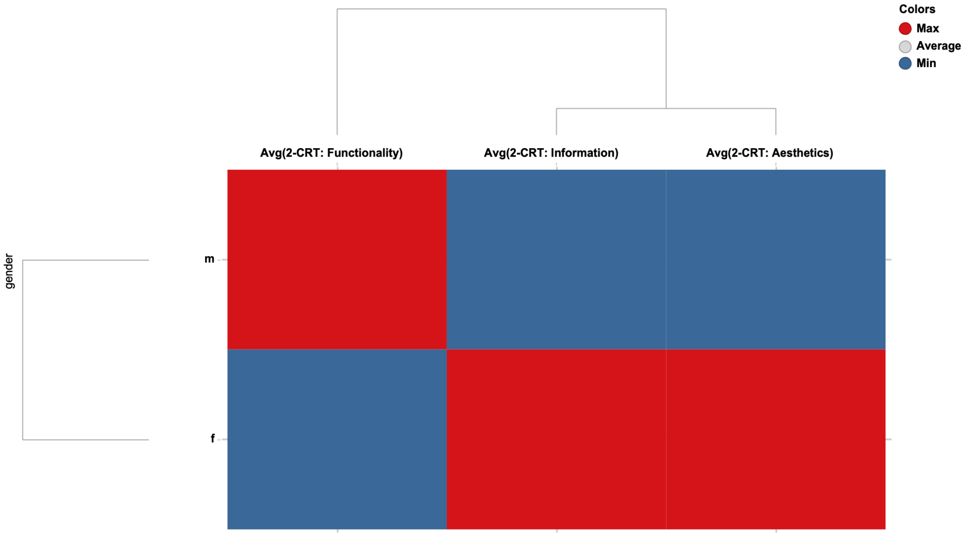


Figure 6.3: Perceived usability of female and male participants in the 2-CRT task (Figure is copied from [34]).

number of awakenings, min awake, step counts, mean RTs in Stroop and 2-CRT tests, and average accuracy in the N-back, Stroop, and 2-CRT tests.

Supervised learning methods were applied to classify bipolar patients and healthy controls. Random Forest (RF) [54], XGBoost [12], Support Vector Machines (SVM) (radial kernel) [79], and K-Nearest Neighbour (KNN) [5] methods were utilised to create their predictive models. Five-fold cross validation was applied to train and test the models. The performance evaluation metrics were accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Area under the Receiver Operating Characteristic Curve (AUC). The latter was used to determine the best performing model. Table 6.5 presents the average of performance metrics calculated for each predictive model. Since XGBoost gave the highest average AUC, the relative variable importance extracted from this tree-based model determined digital phenotypes of individuals' mental health diagnosis. As can be observed in Figure 6.4, **time in bed** is the most significant physiological feature in determining participants' treatment type. **Step count** is the next behavioural feature and **total daily missed counts** as well as **cognitive performance measures of the Stroop test in UbiCAT** are the important cognitive features. It can be inferred that participants' processing speed and executive function as well as their ability to respond during the time limits of the UbiCAT tests were the most important cognitive measures that separate observations of the patients with BD from healthy controls.

Table 6.4: Features collected in a one-week study with UbiCAT (Table is copied from [37]).

#	Name	Category	Type	Features
1	Choice reaction time	Cognitive	Active	Median response time, correct responses, missed stimuli
2	N-back	Cognitive	Active	Mean response time, correct responses, hit rate, false alarm rate, missed stimuli
3	Stroop	Cognitive	Active	Mean response time, correct responses, missed stimuli
4	GPS	Contextual	Passive	Latitudes and longitudes to detect indoor and outdoor environments
5	Time of the day	Contextual	Passive	Time extracted from cognitive test logs
6	Physical Activity	Behavioural	Passive	Step counts, Minutes Sedentary, Minutes Lightly Active, Minutes Fairly Active, Minutes Very Active, Activity Calories
7	Sleep	Physiological	Passive	Minutes Asleep, Minutes Awake, Number of Awakenings, Time in Bed, Minutes REM sleep, light sleep, and deep sleep

Table 6.5: Performance evaluation metrics for classification of healthy and patient groups. (Table is copied from [37]).

Method	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
XGBoost	**77.51±3.28**	76.65±2.91	78.38±4.03	78.60 ±3.83	**76.38±3.45**	**86.40±3.97**
RF	72.63±1.57	69.40±5.32	75.85±3.35	74.65±1.44	71.00 ±2.67	79.60±2.19
KNN	**77.89±4.41**	73.22±10.47	**83.95±3.47**	**83.88±7.18**	71.71±10.85	80.59±1.28
SVM	74.03±3.14	**81.53±4.03**	64.18±7.31	74.13±2.48	73.63±5.31	79.80±5.17

6.3 Chapter Summary

This research showed that UbiCAT as a computing tool for cognitive assessment gives practical test performance measures that were utilised in deriving meaningful insights from subjective human factors and sociodemographics. Moreover, in a one-week study, digital phenotypes of human mental health diagnosis were revealed. Such knowledge are also inline with the literature. Thus, researchers can use cognitive performance measures of the UbiCAT in running their *'in-the-wild'* studies to derive insights from cognitive assessment results, and to analyze the associations between subjective and objective measures of cognition and human factors.

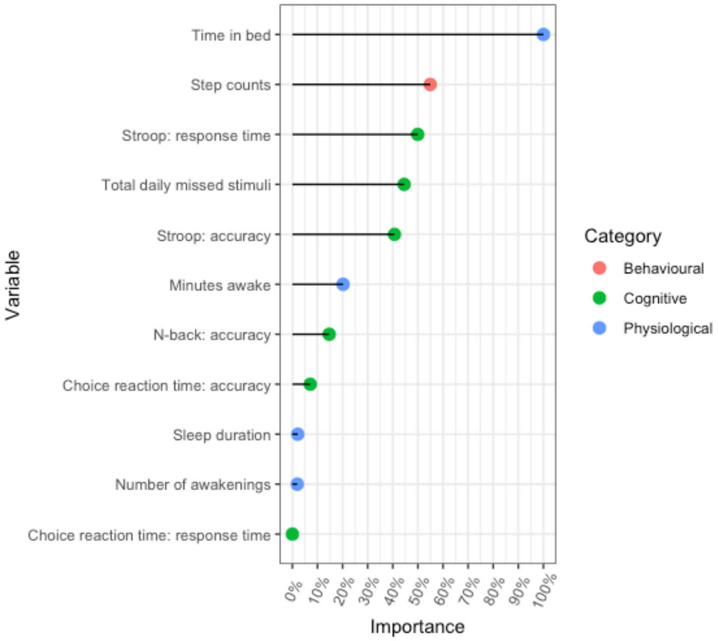


Figure 6.4: Variable importance of the features ranked by XGBoost model (Figure is copied from [37]).

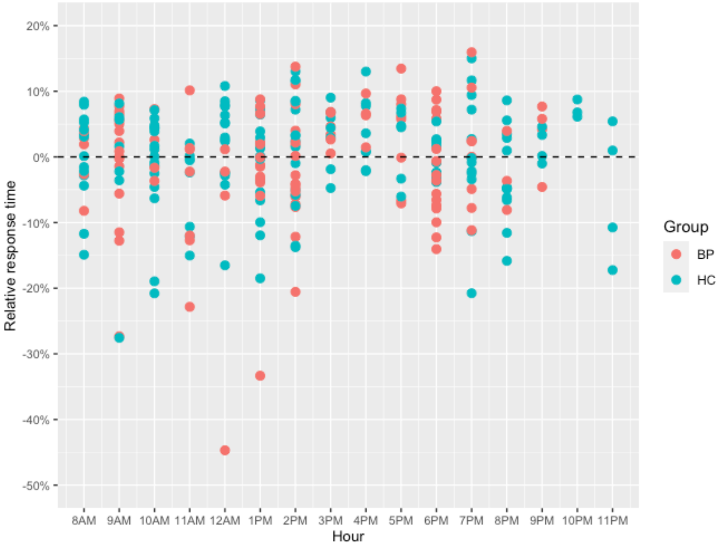


Figure 6.5: Relative alertness of the healthy and patient groups (Figure is copied from [37]).

CHAPTER 7

Discussion

Two pervasive computing technologies were designed and evaluated to assess cognitive functioning: A Web-based cognitive assessment tool called ICAT and a smartwatch-based cognitive assess tool called UbiCAT. Three studies were performed with each of the tools to evaluate their usability, feasibility, and concurrent validity. Usability evaluation of ICAT gave high ratings of psychometric factors by study participants. As such, feasibility of the design and implementation of ICAT was demonstrated as participants stated no specific issue while taking the cognitive tests of ICAT. Concurrent validity of the test scores calculated by ICAT also yielded significant coefficients when compared with gold-standard neuropsychological tests in a population of healthy controls and patients with BD. Likewise, UbiCAT received high usability scores concerning aesthetics, functionality, and information quality and quantity. Followed by that, empirical approaches were used to evaluate feasibility of the design an implementation of UbiCAT. First, an evaluation study revealed significant correlations between the test performance measures of UbiCAT and standard computer-based tools. Second, an *'in-the-wild'* study showed that participants' cognitive test measures were statistically the same in indoor and outdoor places. Such findings provided substantial evidence to demonstrate feasibility of the UbiCAT for *'in-the-wild'* administrations. Significant correlation coefficients between the scores obtained from neuropsychological tests and UbiCAT tests also showed concurrent validity of this tool.

7.1 Computerised Cognitive Assessment Tools

So far, several computerised cognitive test batteries have been built to assess various cognitive domains in patients who suffer from cognitive impairments. Short cognitive test batteries that assess key cognitive functioning of the patients with affective disorders are getting more popular among psychiatrists and psychologists. SCIP is a paper-and-pencil tool that has been validated particularly for patients with affective disorders to provide brief assessments. THINC-it is a patient-administered tool that has acceptable feasibility for assessing cognitive functioning of patients with MDD using short tasks. We utilised such evidence to design and implement ICAT for patients with affective disorders by adapting the cognitive tests from SCIP. Previous related Web-based cognitive assessment tools underwent several clinical studies to evaluate feasibility of their tools and validity of the test scores. Similarly, the studies conducted in this research aimed to evaluate feasibility of the design and implementation of ICAT and concurrent validity of the test scores calculated by this tool. We used speech recognition technology in ICAT for automatic scoring of verbal memory tasks unlike some cognitive test batteries that modified standard verbal memory tasks by asking their users to select the words that they recall from a word list. While a few studies with elderly people used speech recognition technology in their verbal fluency tasks [50, 67, 92, 93], it was not feasible to utilise this technology in a verbal fluency task in ICAT. As such, we replaced this task with another neuropsychological test, which measures the same aspect of cognition.

7.2 Mobile Cognitive Assessment Tools

Functionality of the standard neuropsychological tests have been utilised in mobile apps for smartphones and smartwatches to obtain frequent measures of human cognition unlike lab-based studies, which are schedules once in a while. However, current mobile tools measure limited cognitive domains that give insufficient information about individuals' cognition. In addition, such tools can not collect multivariate sensor data, in particular physiological data (e.g. sleep and heart rate variability). Therefore, UbiCAT was designed to 1) provide short cognitive assessments in majority of contexts and 2) collect wearable and mobile sensor data to identify digital phenotypes of human mental health diagnosis. Previous work evaluated cognitive performance measures of a mobile cognitive assessment tool against computer-based or paper-based neuropsychological tests. While a few tools could not compare some of the cognitive performance measures due to the different test parameters [46, 88], an empirical study conducted in this thesis showed that smartwatch-based and computer-based test measures (e.g. RT and scores) can be compared with each other as long as the test parameters of both platforms are managed to be the same as each other. Two important test parameters are the number of test stimuli per session and the time limit to respond to each stimuli presented in a test session.

A 1-week study on UbiCAT demonstrated feasibility of taking cognitive tests *'in the wild'*. None of the previous related work used GPS data to detect indoor and outdoor places for evaluating feasibility of their cognitive assessment tool. Objective measures of sleep and activity were collected via wearable sensors while previous work often used subjective ratings of sleep and activity measures. Our results showed that sleep, activity, number of missed stimuli in test sessions, and executive functioning are the most important digital phenotypes that classify individuals' mental health diagnosis as bipolar or healthy. While previous work determined digital phenotypes of BD and MDD by analyzing behavioural features using phone sensor data [25, 26, 73, 74], we utilised novel cognitive performance measures of UbiCAT in conjunction with passive mobile and wearable sensor data to identify digital phenotypes with a focus on the cognitive side.

7.3 Limitations

The inherent difficulty in recruiting patients with BD and MDD slightly slowed the progress of our clinical studies. Some of the patients cancelled their follow-up visits right before the meetings and a few of them did not attend their follow-up visits. Consequently, we had to limit our participants or extend the study duration. The lockdown period due to COVID-19 pandemic also forced us to stop recruiting before the study finishes. Despite difficulties in recruiting participants and unpredictable situations, we managed to conduct our clinical studies with patients and healthy controls. In addition, it is not easy to motivate patients with affective disorders to participate in longitudinal studies especially at the time of recruitment. We inferred that the problem is twofold. First, they find it hard to adhere to a schedule every day, in our case taking the UbiCAT tests daily. Second, returning the smartwatch upon finishing the one-week study seemed to be hard for them as they need to schedule it. Although some of the participants were a bit reluctant to take part, they reported that Fitbit smartwatches actually motivated them to be more active, and taking the cognitive tests of the UbiCAT three times per day did not take their time. Two participants could not take part in the feasibility study on UbiCAT due to the constraints imposed by their job requirements.

To deliver a patient-administered tool, we utilised Google's Automatic Speech Recognition (ASR) in the verbal memory tasks of ICAT which resulted in acceptable accuracy. However, existing speech recognition technologies do not have 100% accuracy especially for languages other than English. Thus, there is room for improvement of such technologies. Speaking of verbal memory, we could not implement a standard verbal memory task in UbiCAT since it is considered 'awkward' to repeat a set of words to a smartwatch device especially in public areas. Moreover, limited number of smartwatches are equipped with a microphone, for instance Apple Watch that was excluded from the available options due to its short battery life.

7.4 Future Work

The current pervasive technologies for cognitive assessment are not completely automated to deliver a patient-administered tool. An issue in the existing speech recognition technologies hinders implementing verbal fluency tasks since these technologies often convert the recorded word to the closest in their database. Consequently, users may say irrelevant words and they still receives a positive score. Future research may overcome this deficiency to pave the way for automatic assessment of users' verbal fluency.

A common issue with the cognitive tests, in particular complicated tests involving memory, is that the instruction sets are not read carefully by the users. A study may aim to unobtrusively infer users' comprehension and highlight the parts left unattended to make sure that users are ready to begin the test. Although gaze interaction technologies have paved the way to tackle this issue, it is preferred to avoid using any additional hardware.

As pointed out in previous section, verbal memory tasks have not been implemented for smartwatches. It would be interesting to for future researchers to explore acceptability of deploying automatic verbal memory tasks in wearable devices. In addition, other automated approaches or novel interaction techniques for capturing the recalled words may contribute to the patient-administered tools for cognitive assessments.

So far, there is no usability questionnaire that is particularly designed and validated for smartwatch-based apps. Although smartwatches are considered as mobile devices, we still need to examine specific psychometric factors related to the smartwatch-based apps. In future, researchers can verify validity of their novel usability tool using current smartwatch-based apps including UbiCAT. Perceived comfort in user's dominant hand and the related aspects can be included in future usability questionnaires for smartwatch-based apps.

Digital phenotyping was performed by considering a set of features including activity, sleep, and cognitive test performance measures. Future studies may integrate the cognitive test performance measures of UbiCAT with other data types such as ambient noise and voice features, phone screen time, and phone interaction data including swipes, touches, and typing speed to derive more important features of human mental health diagnosis. Relevant target groups of future studies with an EMA or Experience Sampling Method (ESM) approach on the UbiCAT cognitive tests are ADHD, alcohol drinkers, and drug users.

CHAPTER 8

Conclusion

Internet-based Cognitive Assessment Tool (ICAT) was designed and developed in an interdisciplinary team including computer engineers, psychologists, and psychiatrists to deliver a patient-administered tool adapted from a gold-standard screening tool called Screen for Cognitive Impairment in Psychiatry (SCIP) for affective disorders. Design and implementation of the ICAT were evaluated in a feasibility study, showing that the computerised version of SCIP is feasible and highly usable. To the best of our knowledge, ICAT is the first patient-administered cognitive assessment tool for the patients with affective disorders that utilises speech recognition for automatic assessment of immediate and delayed verbal memory. Moreover, the cognitive tests of ICAT take less than 30 min to complete, which is inline with the intention behind creating short cognitive tests for patients with affective disorders. Consequently, **RQ1** was addressed by showing the successful design and implementation of ICAT for assessing verbal memory, working memory, and psychomotor speed. Such results contribute to the HCI community by informing about the important factors to consider when computerising a paper-and-pencil tool for remote examinations.

Ubiquitous Cognitive Assessment Tool (UbiCAT) is a wearable computing technology for *'in-the-wild'* assessment of three key cognitive functions using smartwatch-based apps. In addition, wearable sensor data collection was utilised in digital phenotyping of human mental health. The cognitive tests of UbiCAT were adapted from neuropsychological tests, and assess attention, working memory, and executive functions. A formative evaluation study on UbiCAT provided preliminary evidence to show that the design of UbiCAT apps works as expected. Followed by that, an empirical evaluation study showed that UbiCAT is highly usable and study participants were generally comfortable when taking the tests on smartwatches as they did not report any significant discomfort in their dominant hand. Thus, **RQ2** was addressed by evaluating participants' perceived psychometric factors in two empirical studies.

Clinical studies were conducted to examine concurrent validity of the test scores calculated by ICAT and UbiCAT against gold-standard neuropsychological tests. Two clinical studies on ICAT revealed significant correlation coefficients between ICAT and SCIP test scores in the working memory, verbal memory, and psychomotor speed tasks as well as total scores. Such findings demonstrate feasibility of ICAT in accurate measurement of cognitive functioning of both healthy persons and patients. Speech recognition technology also yielded acceptable accuracy for both English and Danish languages, demonstrating feasibility of this technology in automatic assessment of immediate and delayed verbal memory of the individuals. Concurrent validity of the UbiCAT test scores was also verified with healthy controls and patients with BD who took neuropsychological tests and UbiCAT cognitive tests at the clinic. We showed that participants' global cognition as well their attention, working memory, and executive functions assessed via UbiCAT correlated significantly with gold-standard neuropsychological tests. Followed by that, participants took part in a one-week study with UbiCAT to take the cognitive tests *in the wild* while their wearable and mobile sensor data were passively collected. An analysis was performed on the cognitive test measures calculated by UbiCAT in indoor and outdoor places, which were detected by analyzing participants' GPS data. Results showed that participants' cognitive performance measures obtained in indoor and outdoor places were statistically the same, indicating feasibility of the UbiCAT cognitive tests. In addition, interviews with the participants revealed that UbiCAT tests were performed during several positions such as sitting, standing, and walking. Such findings demonstrate that smartwatch-based tools are feasible for *ambulatory cognitive assessments.* Taken together, the feasibility of ICAT and UbiCAT were verified through empirical and clinical studies to address **RQ3**.

Unobtrusive wearable and mobile data collection regarding individuals' activity and sleep data assisted in identifying digital phenotypes of healthy and patient groups in conjunction with their daily cognitive performance measures. Supervised learning models of daily cognitive and mobile data showed that participants' time in bed, step counts, daily missed counts during cognitive test session, and measures of executive functions are the digital phenotypes of human mental health diagnosis. Such important features pave the way in building clinical decision support systems for patients with affective disorders to contribute to the early diagnosis and treatments of these patients. We also showed that UbiCAT collects four cognitive test performance measures including the number of correct responses, longest scoring streak, Response Times (RTs), and the number of missed stimuli per test session. The potential of these measures have been demonstrated through the empirical studies on UbiCAT. In particular, the number of missed stimuli was ranked as one of the top features in determining individuals' mental health diagnosis. As such, the Ubicomp community can utilise UbiCAT in their studies with an Ecological Momentary Assessment (EMA) or Experience Sampling Method (ESM) approach to discover further potentials of these objective cognitive performance measures. Moreover, human factors including usability and cognitive load metrics that were evaluated in the empirical studies with UbiCAT discovered unknown associations between subjective and objective factors. An unsupervised approach with a particular focus on individuals' sociodemographics revealed new insights, which inform the HCI community about the potential of user's sociodemographics in personalizing User Interface (UI) of cognitive assessment tools.

Overall, two novel pervasive computing technologies were designed and evaluated to accomplish this thesis. The research questions of this thesis were successfully addressed by building the tools using a User-Centered Design (UCD) approach, and evaluating design, concurrent validity, and feasibility of these tools through empirical and clinical studies.

Part II

Included Papers

CHAPTER 9

Papers Related to ICAT

9.1 Design and Implementation of a Web-based Application to Assess Cognitive Impairment in Affective Disorders

Authors: **Pegah Hafiz**, Kamilla W Miskowiak, Lars V Kessing, Jakob E Bardram

Adjunct Proceedings of the 2018 International Conference on Digital Health.

DOI: https://doi.org/10.1145/3194658.3194691

Design and Implementation of a Web-based Application to Assess Cognitive Impairment in Affective Disorder

Pegah Hafiz
Copenhagen Center for Health Technology, Technical University of Denmark
Lyngby, Denmark
pegh@dtu.dk

Kamilla W. Miskowiak
Psychiatric Center Copenhagen, University Hospital of Copenhagen
Copenhagen, Denmark
kamilla.woznica.miskowiak@regionh.dk

Lars V. Kessing
Psychiatric Center Copenhagen, University Hospital of Copenhagen
Copenhagen, Denmark
lars.vedel.kessing@regionh.dk

Jakob E. Bardram
Copenhagen Center for Health Technology, Technical University of Denmark
Lyngby, Denmark
jakba@dtu.dk

CCS CONCEPTS

- **Software and its engineering** → *Software design engineering*;

KEYWORDS

Design; Implementation; Cognitive Assessment

ACM Reference format:
Pegah Hafiz, Kamilla W. Miskowiak, Lars V. Kessing, and Jakob E. Bardram. 2018. Design and Implementation of a Web-based Application to Assess Cognitive Impairment in Affective Disorder. In *Proceedings of 2018 International Digital Health Conference, Lyon, France, April 23–26, 2018 (DH'18),* 2 pages.
https://doi.org/10.1145/3194658.3194691

1 INTRODUCTION

Affective disorder causes mood disturbance and includes depression and bipolar disorder. Cognitive impairment is one of the determinants of poor functioning in patients suffering from an affective disorder. For example, memory impairment in bipolar patients brings about confusion in their daily life. Other cognitive domains include attention, executive function, and psychomotor speed. Cognitive function of these patients are assessed by means of neuropsychological tests such as California Verbal Learning Test (CVLT) and Trail Making Test (TMT) that are used to examine verbal memory and psychomotor speed, respectively.

The "Screen for Cognitive Impairment in Psychiatry" (SCIP) is a simple and brief screening tool for psychotic disorders including bipolar disorder and depression. It examines cognitive skills namely verbal learning, working memory, verbal fluency, and psychomotor speed [2]. SCIP is a paper-based test battery and is used in clinical setting, in which the examiner explains the instructions and read several words and letter-number sequences to the patient.

However, current computerized test batteries require direct supervision of clinicians in a clinical setting. To our knowledge, none of the computerized test batteries for affective disorders have implemented SCIP in a form of a patient-administered assessment tool.

In this project we are developing a web-based cognitive assessment tool called "Internet-based Cognitive Assessment Tool" (ICAT) for bipolar and depressive patients. This application is a computerized and web-based version of SCIP, in which the third part of SCIP – the verbal fluency task – is replaced with Wechsler Adult Intelligence Scale (WAIS) letter-number sequencing task. In total, ICAT then consists of five sequential tasks, which are explained in section 3.

The aim of this project is to design and implement ICAT as a web-based cognitive assessment tool and examine its validity by running a clinical trial, which compares ICAT with the paper-based SCIP test as the golden standard.

2 RECENT WORKS

Computerized applications for cognitive assessment are currently limited. The Cambridge Neuropsychological Test Automated Battery (CANTAB) [5] is one of the validated test batteries implemented for a wide range of mental disorders. However, CANTAB has inadequate tests to cover affective disorder. The NIH EXAMINER (Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research) [3] is a computerized test battery which measures several cognitive domains. Although this application has multiple tests, a clinician should read a set of words to the patients when assessing verbal memory, which points to the direct supervision of clinicians. THINC-it [4] is a computerized cognitive assessment tool developed for Major Depressive Disorder (MDD) patients. It uses cognitive tasks like Digit Symbol Substitution Test and TMT (part B). However, this system doesn't support cognitive assessment of bipolar patients.

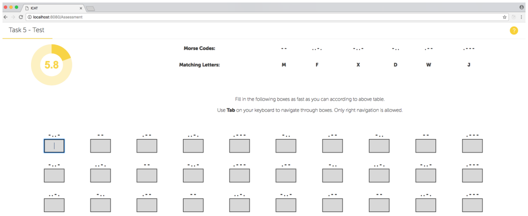
3 ICAT SYSTEM

3.1 Design Methods

The ICAT system is being developed in a user-centered design process involving neuro-psychologists, psychologists, computer

DH'18, April 23–26, 2018, Lyon, France

ACM ISBN 978-1-4503-6493-5/18/04.
https://doi.org/10.1145/3194658.3194691



ICAT

localhost:8080/Assessment

Task 5 - Test

5.8

Morse Codes:	--	..-.	-..-	-..	.--	.---
Matching Letters:	M	F	X	D	W	J

Fill in the following boxes as fast as you can according to above table.

Use **Tab** on your keyboard to navigate through boxes. Only right navigation is allowed.

-..-	--	.--	..-.	.---	-..	.--	-..	--	.---
-..-	..-.	--	-..-	.---	.--	-..	..-.	-..-	.---
..-.	-..	.--	--	..-.	-..-	.--	--	..-.	.---

Figure 1: Task 5 uses a table of Morse codes and their matching letters to assess psycho-motor speed.

scientists, and front-end developers. Once the first prototype of ICAT is implemented, it is been planned to test it with some affective disorder patients.

3.2 ICAT Cognitive Tasks

ICAT includes five tasks which are described below:

- *List Learning*: A list including 10 words are read to the patient using a sound file. The patient should recall as many words as possible. This task is repeated 2 more times with the same set of words. It measures declarative memory and the score ranges from 0 to 30.
- *Consonant Repetition*: It has three letters, a starting number, and a delay (in seconds) for each of the 8 items. The patient counts backwards from the starting number. Then, after a delay, the patient should recall three letters. It measures working memory and the score ranges from 0 to 24.
- *WAIS Letter-Number Sequencing*: It has 7 sets, each set includes 3 letter-number sequences. For each sequence, the numbers should be sorted in ascending order and the letters in alphabetical order. The patient can proceed to the next set only if at least one of the sequences in the current set is reproduced correctly. It measures working memory and the score ranges from 0 to 21.
- *Delayed List Learning*: The word list in task 1 is not played and the patient is asked to recall the earlier words. It measures declarative memory and the score ranges from 0 to 10.
- *Visuo-motor Tracking*: A table including 6 letters and their matching Morse codes are shown to the patient. In 30 seconds, the patient should type the matching letters for 30 Morse codes. It measures executive skills and the score ranges from 0 to 30.(See Figure 1)

3.3 Feedback

Scores of all tasks are displayed to the patients at the end of the assessment. Later during a face-to-face visit, the examiner can interpret the results for the patient and compare his or her performance to a healthy reference group.

3.4 Implementation

The front-end of ICAT system is built using React v16.2.0. We are using the Open m-Health platform [1] as the data back-end. For this purpose, we are designing a Open m-Health JSON schema for cognitive functions, which will be used to store patient’s cognitive profile. This profile includes cognitive skills such as memory and executive function.

In the original paper-based version of the SCIP method, the examiner reads the instructions, words, and letter-number sequences aloud to each participant. The main challenge in the implementation of this system is to convert the role of an examiner from in-person to a digitized format. For this reason, we are examining the use of Google speech recognition web API. It is developed for over 110 languages and will enable us to store each word that patients recall in text format for task 1 and 4.

3.5 Clinical Verification

ICAT will be subject to usability tests focusing on the ability for test subject to understand and perform the cognitive assessment tasks. Once ICAT has been improved based on the usability testing, a clinical verification trial is planned. The goal is to verify and compare the computerized ICAT system against the manual SCIP method as the golden standard.

4 CONCLUSION

We are creating a set of simple and short tasks similar to SCIP in a web application. The use of speech recognition module is supposed to maintain the short duration of the tasks. It has been estimated that 10,000 affective disorder patients in Denmark will use ICAT along with the progress of this PhD project.

5 ACKNOWLEDGMENTS

This project is under development during the early stage of PH’s PhD. PH is an early stage researcher of Technology Enabled Mental Health (TEAM) for Young People. TEAM has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie SkÅĆodowska-Curie grant agreement No. 722561.

REFERENCES

[1] D. Estrin and I. Sim. Open mhealth architecture: an engine for health care innovation. *Science*, 330(6005):759–760, 2010.
[2] J. Gómez-Benito, G. Guilera, Ó. Pino, E. Rojo, R. Tabarés-Seisdedos, G. Safont, A. Martínez-Arán, M. Franco, M. J. Cuesta, B. Crespo-Facorro, et al. The screen for cognitive impairment in psychiatry: diagnostic-specific standardization in psychiatric ill patients. *BMC psychiatry*, 13(1):127, 2013.
[3] J. H. Kramer, D. Mungas, K. L. Possin, K. P. Rankin, A. L. Boxer, H. J. Rosen, A. Bostrom, L. Sinha, A. Berhel, and M. Widmeyer. Nih examiner: conceptualization and development of an executive function battery. *Journal of the international neuropsychological society*, 20(1):11–19, 2014.
[4] R. S. McIntyre, M. W. Best, C. R. Bowie, N. E. Carmona, D. S. Cha, Y. Lee, M. Subramaniapillai, R. B. Mansur, H. Barry, B. T. Baune, et al. The thinc-integrated tool (thinc-it) screening assessment for cognitive dysfunction: Validation in patients with major depressive disorder. *The Journal of clinical psychiatry*, 78(7):873–881, 2017.
[5] T. W. Robbins, M. James, A. M. Owen, B. J. Sahakian, L. McInnes, and P. Rabbitt. Cambridge neuropsychological test automated battery (cantab): a factor analytic study of a large sample of normal elderly volunteers. *Dementia and Geriatric Cognitive Disorders*, 5(5):266–281, 1994.

9.2 The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study

Authors: **Pegah Hafiz**, Kamilla W Miskowiak, Lars V Kessing, Andreas Elleby Jespersen, Kia Obenhausen, Lorant Gulyas, Katarzyna Żukowska, Jakob E Bardram

Published in Journal of Medical Internet Research.

DOI: 10.2196/13898

Original Paper

The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study

Pegah Hafiz[1,2], MSc; Kamilla Woznica Miskowiak[3,4], DMS, DPH; Lars Vedel Kessing[5], MD, DMS; Andreas Elleby Jespersen[3,4], BSc; Kia Obenhausen[3,4], BSc; Lorant Gulyas[2], BSc; Katarzyna Zukowska[2], BEng; Jakob Eyvind Bardram[1,2], MSc, PhD

[1]Digital Health Section, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark

[2]Copenhagen Center for Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark

[3]Department of Psychology, University of Copenhagen, Copenhagen, Denmark

[4]Neurocognition and Emotion in Affective Disorders Group, Copenhagen Affective Disorder Research Centre, Psychiatric Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark

[5]Copenhagen Affective Disorder Research Centre, Psychiatric Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark

Corresponding Author:
Pegah Hafiz, MSc
Digital Health Section
Department of Health Technology
Technical University of Denmark
Richard Petersens Plads
Building 324, 2nd Floor, Room 270
Kongens Lyngby, 2800
Denmark
Phone: 45 91858371
Fax: 45 45253419
Email: pegh@dtu.dk

Abstract

Background: Persistent cognitive impairment is prevalent in unipolar and bipolar disorders and is associated with decreased quality of life and psychosocial dysfunction. The screen for cognitive impairment in psychiatry (SCIP) test is a validated paper-and-pencil instrument for the assessment of cognition in affective disorders. However, there is no digital cognitive screening tool for the brief and accurate assessment of cognitive impairments in this patient group.

Objective: In this paper, we present the design process and feasibility study of the internet-based cognitive assessment tool (ICAT) that is designed based on the cognitive tasks of the SCIP. The aims of this feasibility study were to perform the following tasks among healthy individuals: (1) evaluate the usability of the ICAT, (2) investigate the feasibility of the ICAT as a patient-administered cognitive assessment tool, and (3) examine the performance of automatic speech recognition (ASR) for the assessment of verbal recall.

Methods: The ICAT was developed in a user-centered design process. The cognitive measures of the ICAT were immediate and delayed recall, working memory, and psychomotor speed. Usability and feasibility studies were conducted separately with 2 groups of healthy individuals (N=21 and N=19, respectively). ICAT tests were available in the English and Danish languages. The participants were asked to fill in the post study system usability questionnaire (PSSUQ) upon completing the ICAT test. Verbal recall in the ICAT was assessed using ASR, and the performance evaluation criterion was word error rate (WER). A Pearson 2-tailed correlation analysis significant at the .05 level was applied to investigate the association between the SCIP and ICAT scores.

Results: The overall psychometric factors of PSSUQ for both studies gave scores above 4 (out of 5). The analysis of the feasibility study revealed a moderate to strong correlation between the total scores of the SCIP and ICAT (r=0.63; *P*=.009). There were also moderate to strong correlations between the SCIP and ICAT subtests for immediate verbal recall (r=0.67; *P*=.002) and psychomotor speed (r=0.71; *P*=.001). The associations between the respective subtests for working memory, executive function, and delayed recall, however, were not statistically significant. The corresponding WER for English and Danish responses were 17.8% and 6.3%, respectively.

Conclusions: The ICAT is the first digital screening instrument modified from the SCIP using Web-based technology and ASR. There was good accuracy of the ASR for verbal memory assessment. The moderate correlation between the ICAT and SCIP scores suggests that the ICAT is a valid tool for assessing cognition, although this should be confirmed in a larger study with greater statistical power. Taken together, the ICAT seems to be a valid Web-based cognitive assessment tool that, after some minor modifications and further validation, may be used to screen for cognitive impairment in clinical settings.

(JMIR Form Res 2019;3(3):e13898) doi: 10.2196/13898

KEYWORDS

screening; memory; executive function; bipolar disorder; depression; cognitive impairments; neuropsychological tests; computer software; speech recognition software

Introduction

Background

Cognitive impairment is prevalent in patients with unipolar disorder (UD) and bipolar disorder (BD) even during periods of remission, and it has a negative impact on the quality of life and psychosocial functioning. Nevertheless, cognitive function is rarely assessed in the clinical treatment of these affective disorders because of the time requirement for cognitive tests, which often exceeds the limited health care resources.

To date, there is no patient-administered tool that provides a brief and accurate screening for objective cognitive impairment using gold-standard, performance-based cognitive tasks for patients with affective disorders. The International Society for Bipolar Disorder (ISBD) Targeting Cognition Task Force recently recommended the systematic assessment of cognition in the clinical management of these patients using objective, performance-based cognitive tests [1]. However, validated tests with sensitivity to cognitive impairments in affective disorders only exist in paper-and-pencil or computerized formats, which must be administered by health care professionals. One such test for affective disorders is the screen for cognitive impairment in psychiatry (SCIP). The SCIP is a short (<15 min) paper-and-pencil test administered by trained health care professionals and comprises 5 subtests: (1) list learning (LL), (2) consonant repetition (CR), (3) verbal fluency (VF), (4) delayed list learning (DLL), and (5) visuomotor tracking (VMT) tests. These tests assess verbal recall, working memory, VF, delayed recall, and psychomotor speed, respectively [2]. The ISBD Targeting Cognition Task Force recommends the SCIP for cognitive screening in patients with BD based on recent validation studies [3,4]. In particular, studies point to the validity and reliability of the SCIP for detecting cognitive impairment in BD [5] and UD [6].

Nevertheless, even such brief screening for cognitive impairment in the clinical setting may require too much time and training of health care professionals to be realistic for all patients. This highlights the need for a patient-administered digital tool that provides a brief and valid assessment of cognition with objective cognitive tests, such as the SCIP, for affective disorders.

Previous Studies

Our study is mainly concerned with digital cognitive test batteries, and it partly deals with the application of automatic speech recognition (ASR) in psychiatry. An overview of the related works is presented in the following 2 sections.

Digital Cognitive Test Batteries

In this section, validated digital tools developed for cognitive assessment are presented. Cognitive training tools are, therefore, excluded.

CANTAB Mobile [7] is a validated patient-administered tool to screen for dementia. This app examines memory impairment in patients aged 50 to 90 years using the paired associates learning test. Central nervous system vital signs (CNSVS) is a computerized neurocognitive test battery developed to evaluate cognitive impairment in mental disorders, including UD. The CNSVS has 7 tests, including verbal and visual memory, finger tapping, symbol digit coding, the Stroop test, a shifting attention test, and a continuous performance test [8]. According to the findings by Gualtieri and Johnson, CNSVS is suitable for cognitive assessment and screening of normal subjects. Another test battery is Cogstate, which is aimed to screen patients with Alzheimer's disease but has been used to assess other neuropsychiatric disorders. A recent clinical study on Cogstate [9] aimed to examine cognitive impairment in UD patients compared with healthy controls in terms of psychomotor speed, alertness, visual memory, working memory, verbal memory, and learning and executive functions. Cogstate measures showed impairment in attention and verbal memory and learning, whereas no difference was found in psychomotor speed, visual attention, and working memory in UD patients versus controls. This contrasts with the literature on moderate impairments within these domains in UD and it could be because of the ceiling effects of the Cogstate. The THINC-it is a more recent cognitive assessment tool designed specifically for UD patients that measures attention, working memory, and executive function. This application is the first Web-based patient-administered cognitive screening tool developed for UD and thus represents an important step toward more common assessments of cognition in the clinical management of UD. The THINC-it uses gamified cognitive tasks to engage patients in taking the tests. For example, the *Trails* game is adapted from the trail-making test part B. According to the latest study [10], 100 healthy controls were tested for temporal stability and reliability as well as the validity of the THINC-it. Overall, high stability and reliability and moderate validity were found.

Automatic Speech Recognition in Cognitive Assessment Applications

Recently, ASR has been utilized to examine verbal impairment in mental disorders. Semantic VF as a determinant factor in mild cognitive impairment (MCI) has been automated through

ASR in recent studies [11-14]. Troger et al [14] applied ASR to examine semantic VF in dementia via a telephone-based approach, showing the feasibility of automated analysis in screening for dementia. Toth et al in their recent study [13] derived nonverbal acoustic features, such as the duration of pauses, from ASR among the Hungarian population. Their findings revealed significant differences between healthy individuals and MCI patients in terms of their acoustic features of delayed recall.

The Gaps in the Literature

The limitations of THINC-it are twofold. First, of the cognitive domains assessed by THINC-it, only psychomotor speed shows a moderate correlation with the standardized tests. Second, THINC-it does not examine verbal memory, although this cognitive measure is a predictor of a long-term psychological functional outcome in UD and BD patients [15]. CANTAB Mobile and CNSVS do not assess verbal memory as well. The lack of verbal memory assessment might be partially because of uncertainty about how to measure it via a digital tool. Moreover, the tests suggested by CANTAB and CNSVS for use in affective disorders have not been specifically developed to screen for cognitive impairment in UD and BD patients and may thus not have optimal sensitivity for impairments in these groups.

Internet-Based Cognitive Assessment Tool

We developed the internet-based cognitive assessment tool (ICAT) with the perspective that it can be administered by patients themselves at home. Specifically, the ICAT is a Web-based cognitive test battery that examines immediate and delayed verbal recall, working memory, executive function, and psychomotor speed in 5 short tasks. Speech recognition technology has become advanced enough to be used in various applications. Moreover, ASR requires minimum technology and resources for remote examination. Therefore, ASR is utilized in 2 ICAT subtests to assess immediate and delayed verbal recall.

Goals of This Study

The objective of this paper is threefold: first, to present the ICAT as a Web-based cognitive test battery designed based on the cognitive tests included in the SCIP; second, to present 2 studies assessing 2 aspects of the ICAT—(1) its usability and (2) its feasibility evaluated by correlation analysis between the SCIP and ICAT subtests and total scores; third, to evaluate the accuracy of the ASR for immediate and delayed verbal recall.

Methods

Design Methods

The ICAT user interface (UI) was designed in a user-centered design process involving computer scientists, health informaticians, psychiatrists, and psychologists. Overall, the design process took 5 months and was performed in 4 consecutive stages, as explained below.

Phase 1: Brainstorming Design Sessions

The essential components of the ICAT system as a patient-administered system were brainstormed in 3 weekly meetings. In addition, the technical opportunities and limitations of computerizing the SCIP subtests were investigated.

Phase 2: Personas and User Interface Design

To identify design requirements and system functionalities, 2 personas were prepared based on the inputs received from psychiatrists and psychologists, who provided the practical lived experiences of the patients. A flowchart was created based on the personas to determine the navigation through different components (eg, homepage, instructions, and cognitive assessment tasks), and UI wireframes of each page were drawn.

Phase 3: Mock-Up

The wireframes were presented as a slideshow and thoroughly discussed by the ICAT team members during user experience (UX) prototyping sessions. During these sessions every aspect of the ICAT was (re)designed, including the layout and graphical design of each page, the instructions, the use of speech recognition, the feedback to the users, the use of input modalities (ie, keyboard and mouse), and the informed consent pages. During the design process, the original SCIP tasks were significantly modified for administration on Web-based technology in a browser, particularly considering support for a PC-based setup with keyboard and mouse. In this phase, the homepage of the ICAT contained a welcome page and a speaker test (see Multimedia Appendix 1).

Phase 4: Prototyping

The low-fidelity mock-up of the ICAT was gradually turned into a functional prototype using Web technology for graphical rendering in a browser but with no storage or persistence. This prototype was used for the initial assessment during UX prototyping sessions involving PH, KWM, LVK, and JEB. The slideshows created during phase 3 were expanded to 4 pages in the low-fidelity mock-ups (see Multimedia Appendix 1); the first page was added to determine how the patient would be notified to take the test, and the fourth page was the consent form. The final prototype was used to deploy the ICAT application on a Web platform.

System Description

The ICAT includes the following 3 overall sections, which are presented one after another to the user: (1) the homepage, including an introduction, general instructions, and an informed consent form; (2) the technical setup (speaker and microphone test), and (3) cognitive assessment tasks. The ICAT supports both English and Danish, and users can hence select their preferred (native) language before proceeding to the general instructions. For readability, the lengthy instructions were divided into multiple pages. The terms of use in the consent form clarifies the purpose of the study, what data are gathered, and how the user’s data will be handled. All of this complies with the European data protection law (general data protection regulation, GDPR). As the ICAT makes extensive use of ASR, the second section (technical setup) ensures that the microphone and speakers are properly configured. See Multimedia Appendix

1 to check the final design of the ICAT homepage and technical setup, respectively. The third section of the ICAT contains a set of 5 short tasks, each including a test introduction and task-specific instructions. These 5 tasks were modified versions of the following:

- SCIP LL
- SCIP CR
- Wechsler Adult Intelligence Scale letter-number sequencing (WAIS LNS)
- SCIP DLL
- SCIP VMT

All of the ICAT subtests were adapted from the SCIP except for the third subtest that was replaced with a modified version of WAIS LNS. A detailed description of each ICAT task can be found in Table 1. The ICAT WAIS LNS and VMT subtests present a practice set to the users beforehand. The practice sets were adapted from their corresponding clinically administered tests. In total, the 5 tasks of the ICAT take 20 to 30 min to complete.

Table 1. Description of the internet-based cognitive assessment tool subtests.

Task features	Task 1: list learning[a]	Task 2: consonant repetition[b]	Task 3: Wechsler Adult Intelligence Scale letter-number sequencing[c]	Task 4: delayed list learning[d]	Task 5: visuomotor tracking[e]
Measure	Verbal memory (immediate recall)	Working memory	Working memory	Delayed verbal memory (delayed recall)	Psychomotor speed
Scoring criteria	Total number of correctly recalled words for 3 trials	Total number of correctly recalled letters	Total number of correctly sorted sequences	Total number of correctly recalled words	Total number of correct matching letters
Score range	0–30	0–24	0–21	0–10	0–30
Practice test	No	No	Yes	No	Yes

[a] An audio file containing a list of 10 words is played to the patient. Following that, the patient recalls as many words as possible and speak them aloud. This task is repeated 2 more times (3 trials in total) using the same word list.

[b] First, a sequence of letters is played via an audio file. Then, the patient should sort a set of numbers in descending order within a certain time period (this task is only for delaying the response). After time is up, the patient recalls and types the letters that were read to him or her earlier.

[c] A set of letter-number sequences are displayed on the screen one by one. Each sequence is played via an audio file to the patient. Following that, the patient sorts the numbers and letters of the sequence and types them.

[d] In this task, the patient should recall the same words that were played in the first list learning task and speak them aloud. No audio is played for the patient in this task.

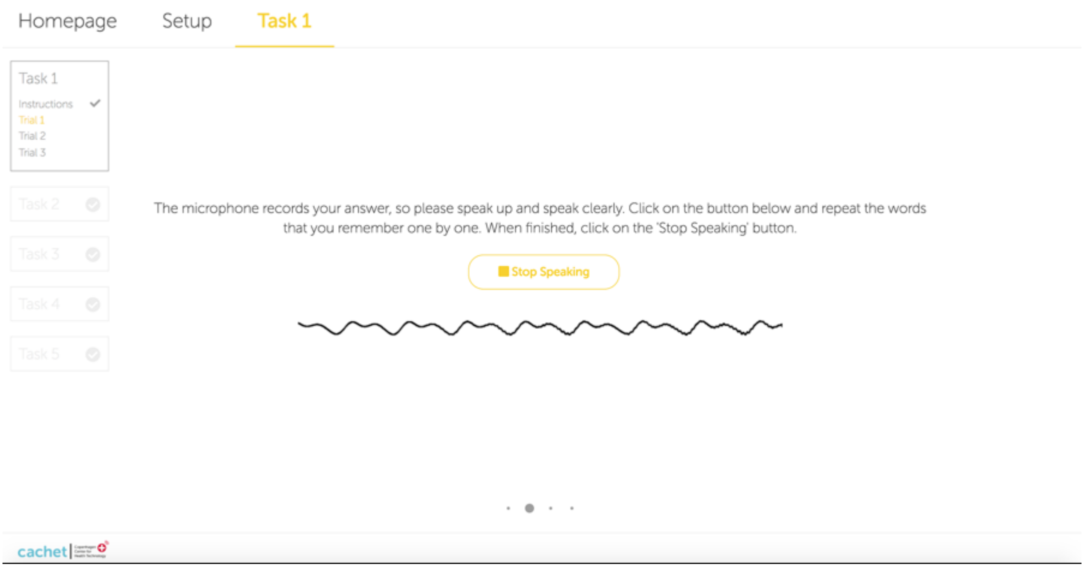
[e] A table including 6 letters and their matching codes (a combination of circles and asterisks) is shown to the patient. In 30 seconds, the patient enters the matching letters of 30 random codes one by one.

Modified Elements of the Screen for Cognitive Impairment in Psychiatry Tasks

List Learning and Delayed List Learning Tasks: Utilizing Speech Recognition

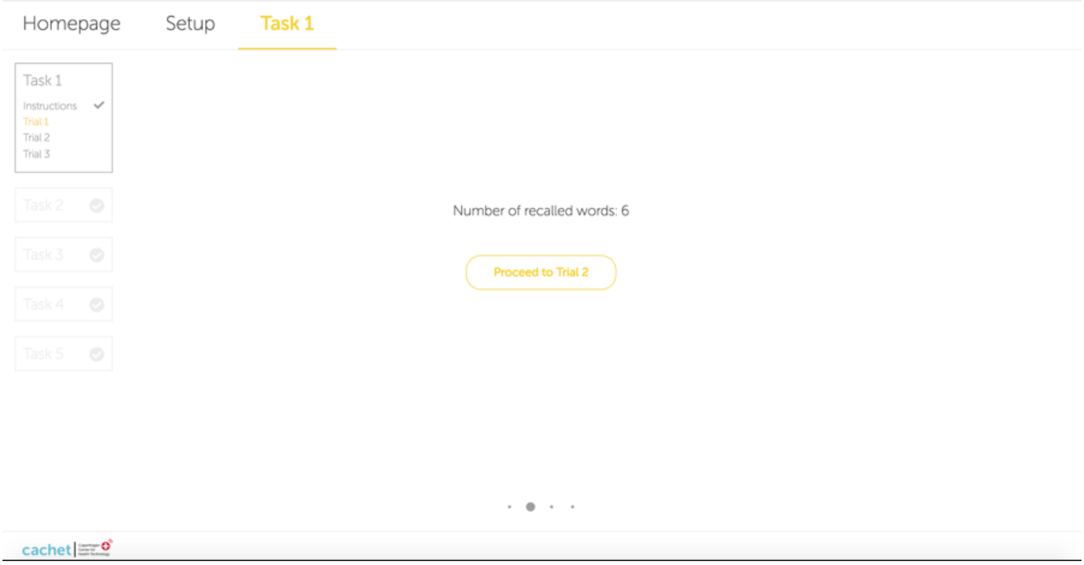
During the initial design of the ICAT LL and DLL subtests, users were supposed to type the recalled words. However, typing was not a suitable input technique for 3 reasons. First, typing influences human visual short-term memory that may help the users in practicing the words. Hence, practicing could significantly increase the users' scores in the second and third trials of the LL task. Second, typing skill depends on the people's age and previous typing experience. Third, misspelled words may cause a problem when scores are automatically calculated. To clarify the latter, the SCIP administrator reads the words aloud and gives scores based on what he or she hears from the patient. Hence, giving a score to a misspelled word is unclear. An editing option for the ASR transcript could allow users to check and modify it after a recall phase. However, this approach would display the words to the users, which would then significantly improve their verbal scores because (1) all trials of the LL subtest use the same set of words and (2) it would not comply with the SCIP administration manual. By considering these major issues, the alternative to typing was to utilize ASR. Figure 1 shows the UI of the ICAT LL subtest including a user's sound wave received from the microphone device during a recall phase. Figure 2 displays the number of recalled words calculated based on the real-time ASR when the user stops speaking. The ICAT DLL task has an interface and functionality similar to the LL task except that no audio file is played for the users.

Figure 1. Screenshot of a sound wave received from a user's microphone device during a recall phase of the internet-based cognitive assessment tool list learning task.



The microphone records your answer, so please speak up and speak clearly. Click on the button below and repeat the words that you remember one by one. When finished, click on the 'Stop Speaking' button.

Figure 2. Screenshot of the number of recalled words recognized by automatic speech recognition when the user stops speaking in the list learning task.

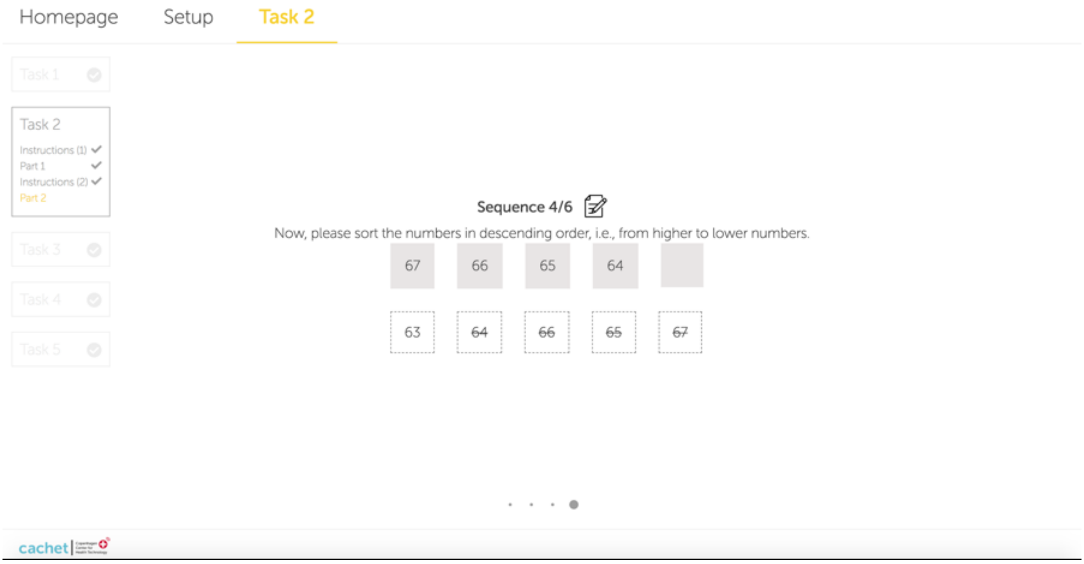


Number of recalled words: 6

Consonant Repetition Task: Sorting Numbers Using Drag-and-Drop

During the SCIP CR task, the test administrator asks the patient to count backwards by starting from a specific number for a time period. We replaced this face-to-face countdown with a sorting module in the ICAT CR subtest, where the users should drag each number and drop it into its correct place. The numbers displayed on the user's screen should be placed in descending order. Figure 3 shows a sample drag-and-drop task where users should sort a sequence of numbers from 67 (highest) to 63 (lowest) within a certain time limit. Each sequence includes 5 numbers, and if the user sorts them correctly, the next set automatically appears on the screen.

Figure 3. Screenshot of the internet-based cognitive assessment tool consonant repetition task where the user should sort the numbers in descending order by dragging and dropping the numbers into their correct place.



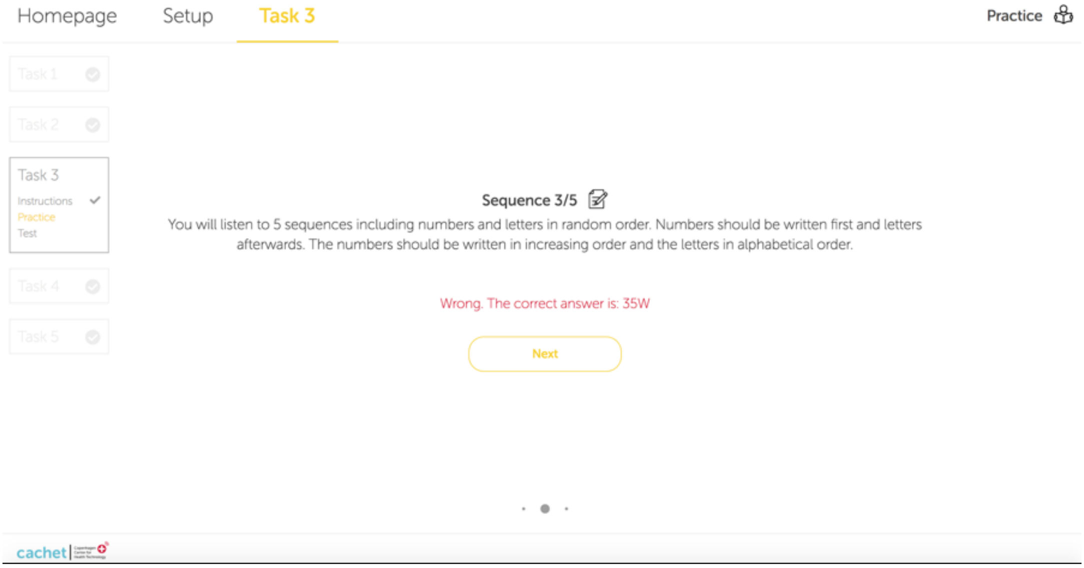
Sequence 4/6

Now, please sort the numbers in descending order, i.e., from higher to lower numbers.

Wechsler Adult Intelligence Scale Letter-Number Sequencing Task: Replacing Verbal Fluency

The SCIP subtest for the assessment of the VF requires the patient to generate as many words as possible that start with a specific letter, for example, *F* in 30 seconds. The third subtest of the ICAT uses WAIS LNS because the SCIP VF task could not be implemented adequately in the technology. Hence, VF was replaced with WAIS LNS, which measures executive function. Figure 4 shows an example of an incorrect response to a stimulus during a practice test of the ICAT WAIS LNS subtest.

Figure 4. The internet-based cognitive assessment tool Wechsler Adult Intelligence Scale letter-number sequencing task includes a practice set with 5 sequences to prepare the user for the actual test. This screenshot shows that a user sorted a sequence incorrectly.

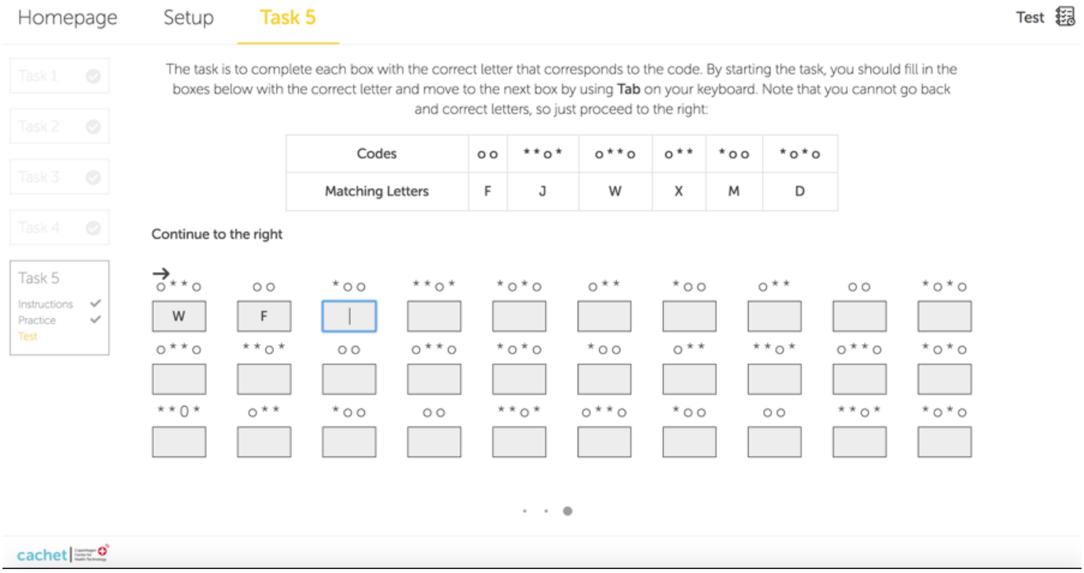


Sequence 3/5

You will listen to 5 sequences including numbers and letters in random order. Numbers should be written first and letters afterwards. The numbers should be written in increasing order and the letters in alphabetical order.

Wrong. The correct answer is: 35W

Figure 5. The user interface of the internet-based cognitive assessment tool visuomotor tracking task, where the user should enter the matching letter for each symbol as fast as possible.



The task is to complete each box with the correct letter that corresponds to the code. By starting the task, you should fill in the boxes below with the correct letter and move to the next box by using **Tab** on your keyboard. Note that you cannot go back and correct letters, so just proceed to the right:

Codes	oo	**o*	o**o	o**	*oo	*o*o
Matching Letters	F	J	W	X	M	D

Continue to the right

→

o**o	oo	*oo	**o*	*o*o	o**	*oo	o**	oo	*o*o
W	F	|							
o**o	**o*	oo	o**o	*o*o	*oo	o**	**o*	o**o	*o*o
**0*	o**	*oo	oo	**o*	o**o	*oo	oo	**o*	*o*o

Visuomotor Tracking Task: Changing Morse Codes

A table of 6 letters and their corresponding codes is written for the patients during the test, and they are required to write down the matching Morse code of 30 letters on a paper within 30 seconds. Owing to slow typing, especially among elderly people, we decided to ask users to enter the matching letter of each code in the ICAT VMT task. The Morse codes of the SCIP VMT task were modified from dots and dashes into a different combination of circles and asterisk symbols because of the learning effect for those participants who are already familiar with the Morse codes. The earlier design of this task can be found in our previous publication [16]. According to the former design of this task, a countdown clock was displayed to the user during the test, but it was later removed to prevent distraction. Figure 5 shows the current design of the ICAT VMT task.

Technical Specifications and Apparatus

The low-fidelity mock-up of the ICAT was created in the Balsamiq desktop app [17]. The front end of the ICAT was built using React (version 15.4.0) developed by Facebook incorporation company. The Copenhagen Center for Health Technology—CACHET Research Platform (CARP), which implemented an open mobile health (mHealth) data storage unit [18], was used as the data back end, and ICAT-specific JavaScript object notation (JSON) schemas for the cognitive functions were designed according to the open mHealth specifications. Google's ASR service [19] was used in the LL and DLL subtests, which require Google Chrome to run the application. CARP and the ICAT system are deployed on secure servers at the Technical University of Denmark. For the evaluation and feasibility studies, ICAT tests were administered using a MacBook Pro (Retina 15 inch) laptop and an external mouse for those who were not comfortable with the MacBook touchpad. Pearson correlation analysis was performed in SPSS.

Usability and Feasibility Studies

The local ethics committee for the Mental Health Services, Capital Region of Denmark, determined that their permission for the study was not needed because it involved no testing of biomedical products nor involved any invasive procedures. A total of 2 studies were conducted: the first study was a usability test, which we will refer to as Study 1, and the second is a feasibility study, which will be called Study 2 in the rest of this paper. Participants of both studies signed an informed consent before the data collection. The informed consent was compliant with the GDPR regulation to protect the personal data of the users. In the following sections, we elaborate on the participants and procedures of the studies individually.

Participants

All participants were healthy individuals. Study 1 included healthy students and individuals from the campus of the Technical University of Denmark and the city of Copenhagen. The inclusion criterion was English or Danish language skills, and the exclusion criterion was any hearing disability because some of the ICAT tasks used audio files. Study 2 included healthy participants who were recruited from blood banks at hospitals within the Capital Region.

Procedure

The age and gender of the subjects were collected before conducting both studies. Study 1 was conducted during June and August 2018. The study leader (PH) first asked the native language of the participant. Then, PH introduced the ICAT system to the participant and briefly explained the purpose of the study. The think-aloud method [20] was used during the test. The participants were not supposed to receive assistance during the test except for login issues. Study 2 was conducted during August and September 2018. Each participant first performed the Danish version of the SCIP (SCIP-D) as administered by research assistants in the Neurocognition and Emotion in Affective Disorders group (AEJ, KO) and then completed the ICAT test.

The usability of the ICAT UI was evaluated in both studies by the poststudy system usability questionnaire (PSSUQ) [21]. Upon completing the ICAT test, the PSSUQ questionnaire was sent to the subjects' email via Google Form, and the study leaders conducted a brief follow-up interview with the participant. During the interview, the participants were asked to mention any general or task-specific issues or suggestions. Participants could also type further comments at the end of the PSSUQ form. The voice of the users was recorded during the ICAT test and the follow-up interviews. The manually generated transcripts of the participants' verbal responses during the ICAT LL and DLL subtests were obtained from their recorded files.

Metrics

Usability Factors

PSSUQ includes 19 items, each rated on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). The psychometric factors of the PSSUQ are (1) overall usability, (2) system usefulness, (3) information quality, and (4) interface quality.

Word Error Rate

Previous studies used word error rate (WER) as the performance measure of ASR [11,12,14]. If N is the total number of words, D is the number of deletions, S is the number of substitutions, and I is the number of insertions, then, WER = (S+D+I)/N. WER is calculated by comparing ASR transcripts to the manually generated transcripts for English and Danish responses during the ICAT LL and DLL subtests.

Correlation Analysis

Pearson 2-tailed correlation analysis was performed at the .05 significance level for both the SCIP and ICAT subscores and total scores of the participants of Study 2.

Data Exclusion

The ICAT data of the WAIS LNS subtest were lost for 3 participants of Study 2. The correlation analysis was, therefore, performed for 16 participants.

Results

User Statistics

Study 1 included N=21 subjects—9 females and 12 males, with an average age of 31 years (SD 12). Of the Danish-speaking participants, 7 were native Danish speakers and 2 were citizens of Copenhagen who had spoken Danish for at least 10 years. As for the rest of the participants, 1 was a native English speaker and 11 spoke other languages. Study 2 included N=19 subjects—13 females and 6 males, with an average age of 36 years (SD 15). All participants of this study had Danish as their native language.

Internet-Based Cognitive Assessment Tool Test Scores

The scores obtained by the participants of both studies in tasks 1-5 are shown in Figures 6-10.

Figure 6. Boxplots of the internet-based cognitive assessment tool and screen for cognitive impairment in psychiatry subscores of the participants of both studies in task 1.

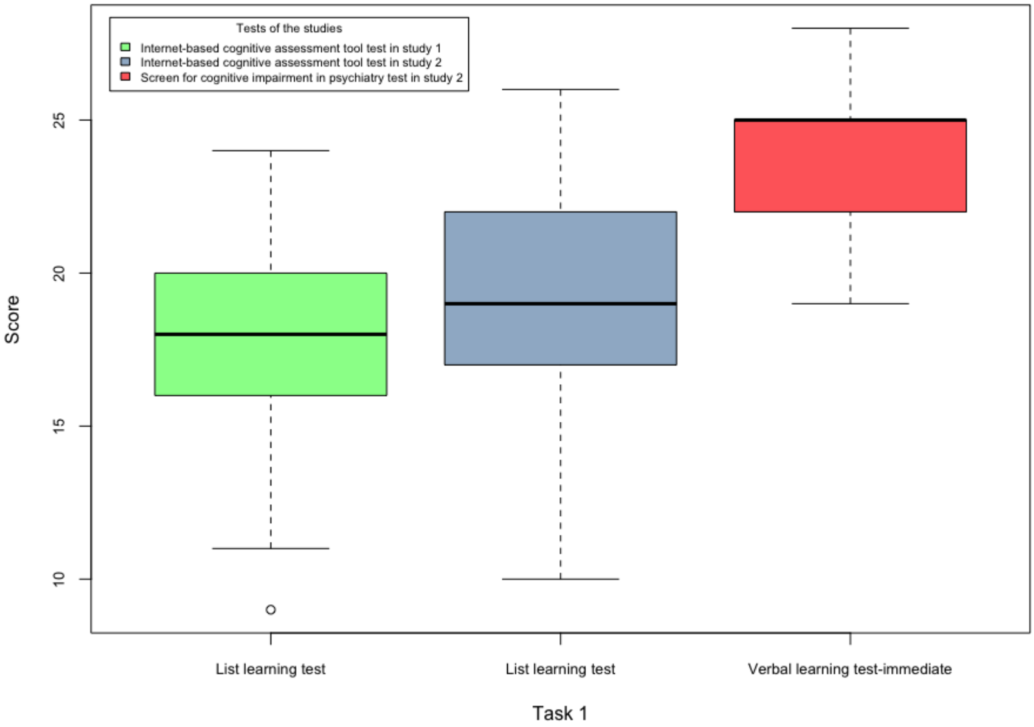


Figure 7. Boxplots of the internet-based cognitive assessment tool and screen for cognitive impairment in psychiatry subscores of the participants of both studies in task 2.

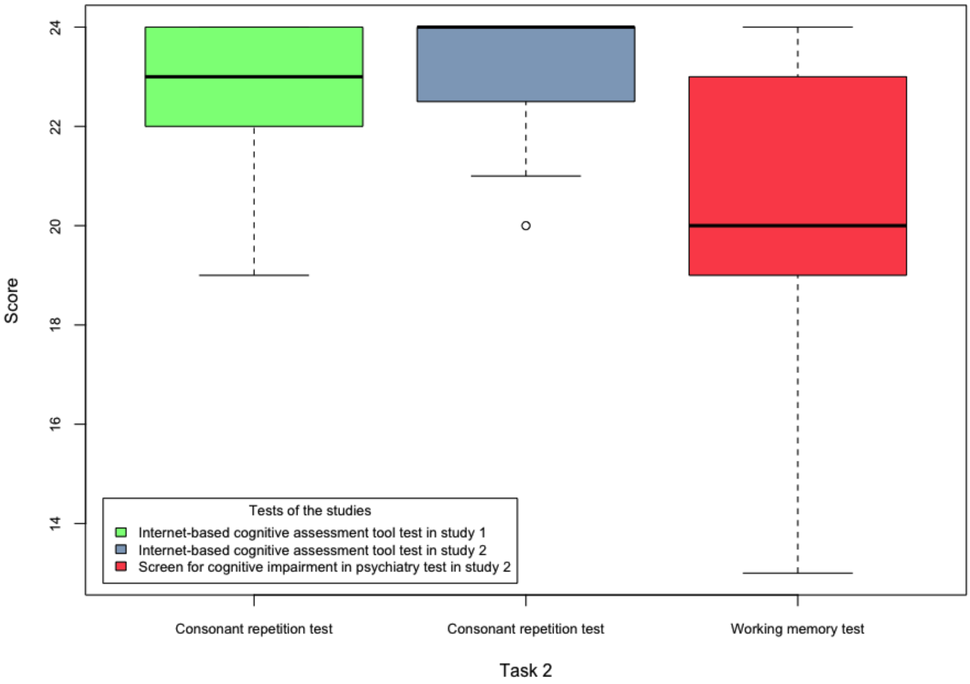


Figure 8. Boxplots of the internet-based cognitive assessment tool and screen for cognitive impairment in psychiatry subscores of the participants of both studies in task 3.

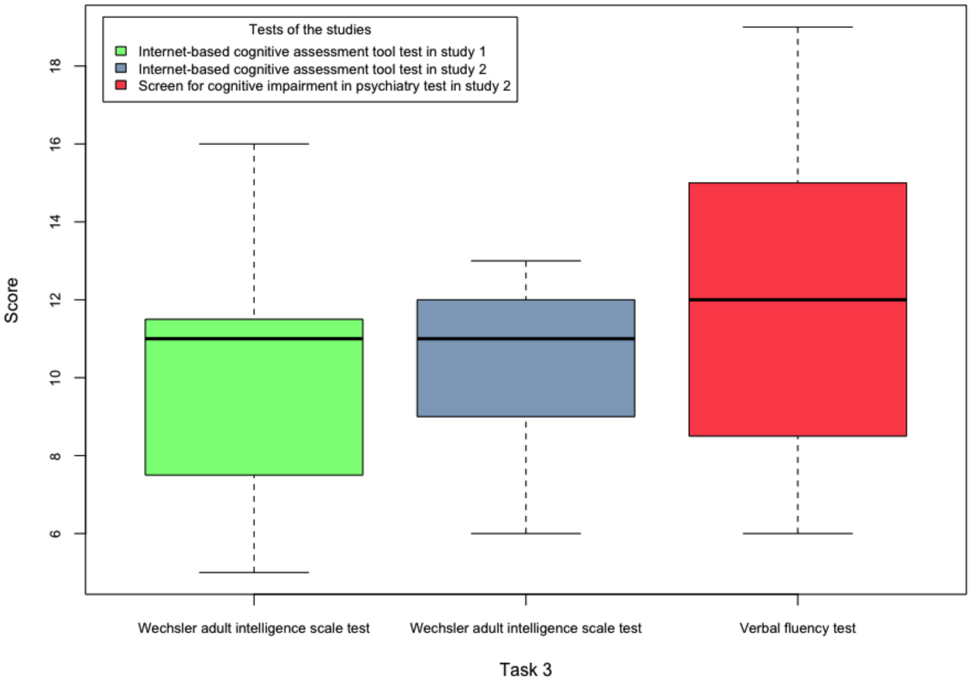


Figure 9. Boxplots of the internet-based cognitive assessment tool and screen for cognitive impairment in psychiatry subscores of the participants of both studies in task 4.

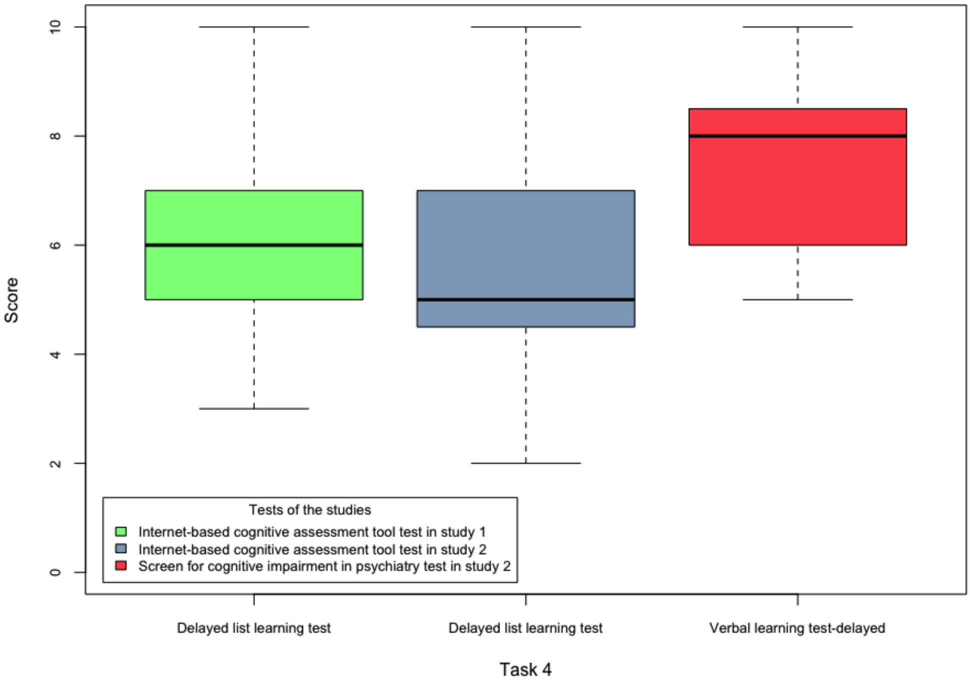
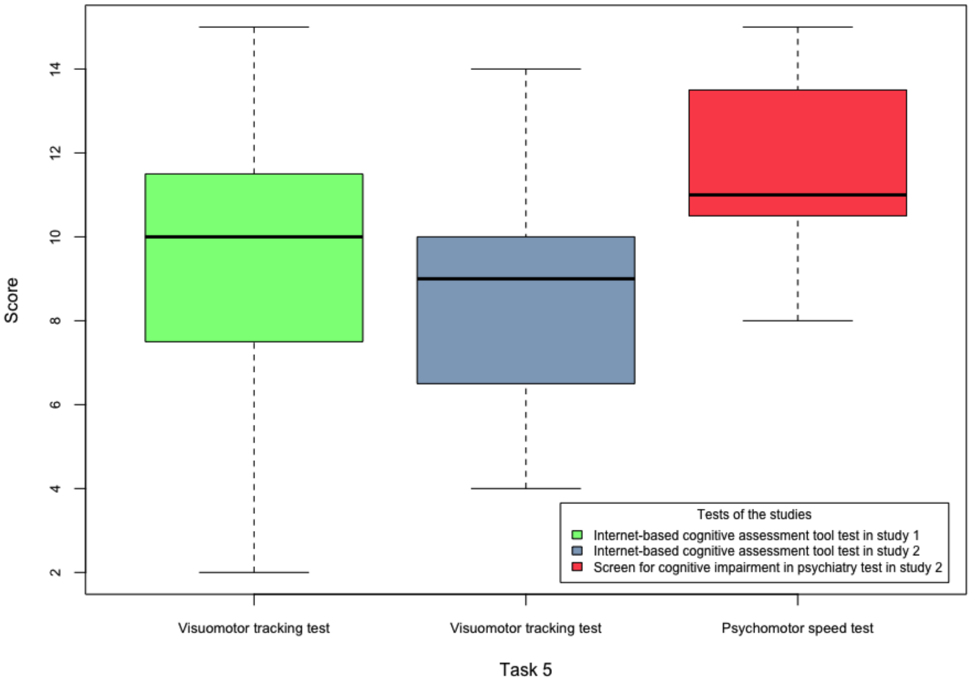


Figure 10. Boxplots of the internet-based cognitive assessment tool and screen for cognitive impairment in psychiatry subscores of the participants of both studies in task 5.



Usability and Feasibility Outcomes

Of the total number of subjects in both studies (N=40), 37 participants submitted the PSSUQ. The psychometric factors of the PSSUQ results (Table 2) are reported for each study separately because the objectives and procedures of those studies were different. Moreover, the PSSUQ results are calculated for Danish and English test participants. According to the reports collected from the follow-up interview and additional comments received via the PSSUQ form, some of the participants reported some issues and gave some suggestions regarding the instructions and the functionality of the ICAT tests. A total of 2 participants of Study 1 mentioned that there were too many instructions in the ICAT LL task. A participant of Study 1 said that the sorting module in the ICAT CR task was complicated and thus not user friendly, and 2 participants of Study 1 mentioned that this module was problematic. In total, 2 participants of Study 2 mentioned that the ICAT CR task was far easier than the SCIP CR task. A participant of Study 1 suggested replacing some of the textual information in the instructions of the ICAT WAIS LNS task with an example. We did not receive any comment on the ICAT DLL task, perhaps because its functionality was similar to the ICAT LL task. For the ICAT VMT task, a participant of Study 1 mentioned that the time limit of this task was too short. A total of 2 participants of Study 1 mentioned that the practice sets of the ICAT CR, WAIS LNS, and VMT were helpful in understanding the tests.

The results of the correlation analysis between the SCIP-D and ICAT subscores and total scores can be found in Table 3.

The analysis of ASR for the ICAT LL and DLL tasks are reported in Table 4. As can be seen, the insertion (I) rate is 0 for both languages. The number of recalls versus recognition accuracy of each English and Danish word are represented in Figures 11 and 12, respectively. Overall, 332 words were received from 12 English-speaking participants of Study 1 and 887 words were gathered from 28 Danish-speaking subjects (9 from Study 1 and 19 from Study 2). Note that the words which are repeated more than once are included in Figures 11 and 12. Of the English words, *machine*, *milk*, and *coffee* were the most recalled and the least misinterpreted words, whereas *bed* and *hat* were highly misinterpreted and were the least memorized terms. The word *garden* was the most recalled word (45 times) but its accuracy (77.78%) was not as high as the words mentioned earlier. For the Danish word list, *mælk* and *sømand* were correctly recognized for every response received, whereas *seng* and *brev* were misinterpreted frequently.

Table 2. Psychometric factors of poststudy system usability questionnaire for usability evaluation of the internet-based cognitive assessment tool reported for both studies and testing languages.

Factor	Study 1 (N=21), mean (SD)	Study 2 (N=16), mean (SD)	Danish test (N=25), mean (SD)	English test (N=12), mean (SD)
Overall score	4.12 (0.46)	4.36 (0.42)	4.25 (0.45)	4.19 (0.45)
System usage	4.23 (0.53)	4.52 (0.41)	4.39 (0.48)	4.35 (0.45)
Information quality	3.86 (0.55)	4.24 (0.58)	4.11 (0.55)	3.84 (0.64)
Interface quality	4.28 (0.62)	4.25 (0.49)	4.16 (0.57)	4.50 (0.45)

Table 3. Results of correlation analysis applied to the screen for cognitive impairment in psychiatry (Danish version) and internet-based cognitive assessment tool scores.

Cognitive domain	Screen for cognitive impairment in psychiatry–Danish version task	Internet-based cognitive assessment tool task	Pearson correlation coefficient (r)	*P* value
Verbal learning (SCIP-2[a])—using ASR[b] transcripts	VLT[c]-I	LL[d]	0.56	.013
Verbal learning (SCIP-3[e])—using ASR transcripts	VLT-I	LL	0.67	.002
Verbal learning (SCIP-3)—using manual transcripts	VLT-I	LL	0.66	.002
Working memory (SCIP-2)	WMT[f]	CR[g]	−0.12	.63
Working memory (SCIP-3)	WMT	CR	0.11	.65
Executive function (SCIP-3)	Verbal fluency test	Wechsler adult intelligence letter-number sequencing	0.29	.27
Delayed recall (SCIP-3)—using ASR transcripts	VLT-D[h]	DLL[i]	0.34	.15
Delayed recall (SCIP-3)—using manual transcripts	VLT-D	DLL	0.58	.009
Psychomotor speed (SCIP-3)	VMT[j]	VMT	0.71	.001
Total score	Total	Total	0.63	.009

[a] SCIP-2: Screen for Cognitive Impairment in Psychiatry–version 2.

[b] ASR: automatic speech recognition.

[c] VLT-I: verbal learning test-immediate.

[d] LL: list learning.

[e] SCIP-3: Screen for Cognitive Impairment in Psychiatry–version 3.

[f] WMT: working memory test.

[g] CR: Consonant Repetition.

[h] VLT-D: verbal learning test–delayed.

[i] DLL: delayed list learning.

[j] VMT: visuomotor tracking.

Table 4. Performance evaluation of automatic speech recognition in internet-based cognitive assessment tool task 1 (list learning) and task 4 (delayed list learning).

Language	Participants in task 1, n	Participants in task 4, n	Average word error rate	Substitution error ratio, %	Deletion error ratio, %
English	12	11[a]	17.77	77.97	22.03
Danish	28	27[b]	6.31	92.98	7.02

[a] 1 English-speaking participant accidentally clicked on the stop button in the internet-based cognitive assessment tool delayed list learning task before repeating the recalled words.

[b] 1 Danish participant could not remember any word in the internet-based cognitive assessment tool delayed list learning task.

Figure 11. Total number of recalls versus the recognition accuracy of the English words in task 1 (list learning) and task 4 (delayed list learning).

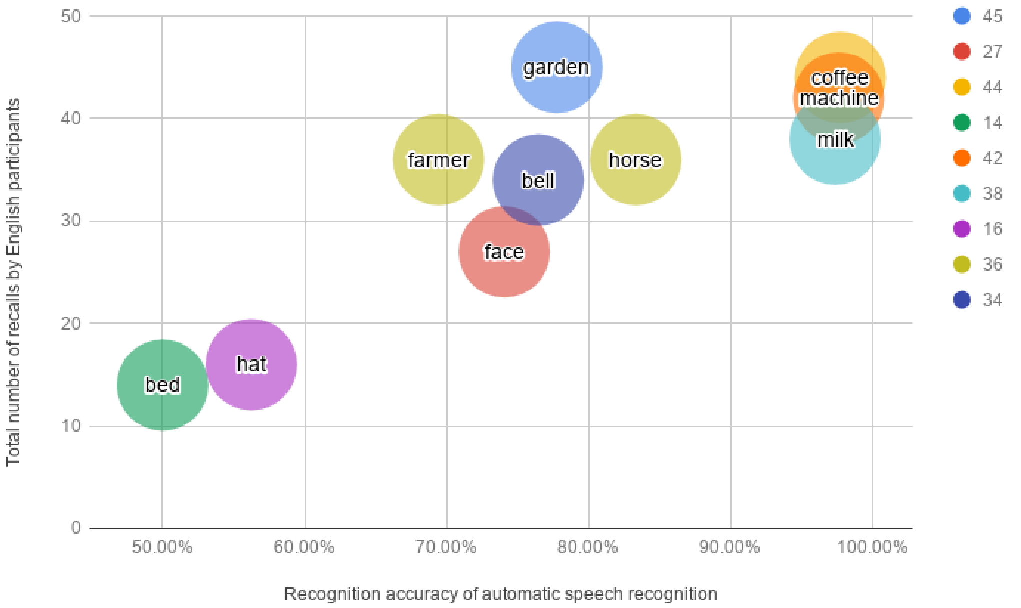
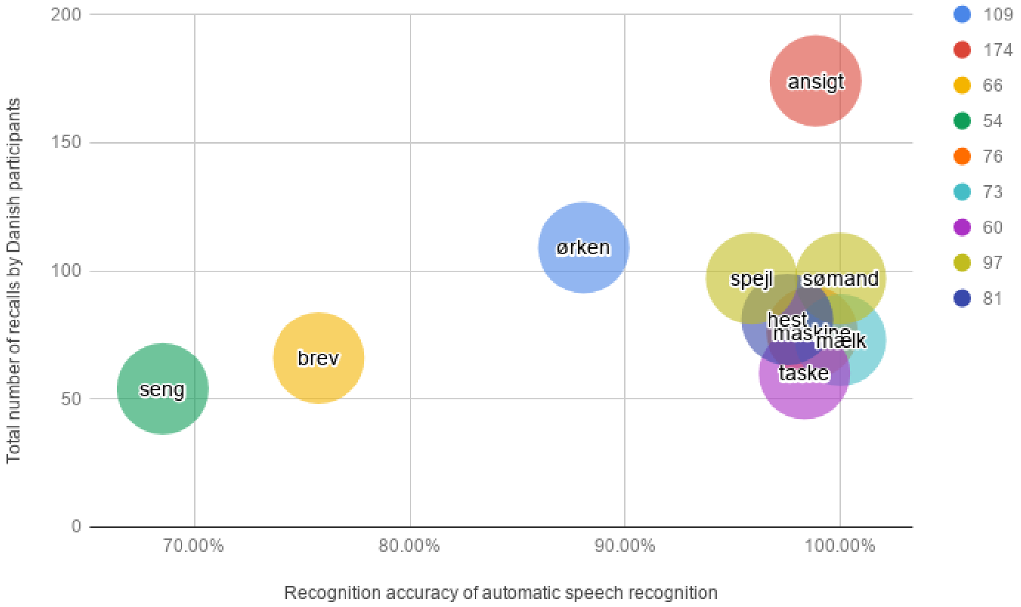


Figure 12. Total number of recalls versus the recognition accuracy of the Danish words in task 1 (list learning) and task 4 (delayed list learning).



Discussion

Principal Findings

The ICAT is the first Web-based cognitive screening tool for affective disorders, designed based on the SCIP as a gold-standard tool, and it uses ASR to assess immediate and delayed verbal recall. The key findings were that the ICAT was easy to use, had promising feasibility outcomes in measuring key cognitive functions, and had acceptable concurrent validity. Specifically, the ICAT and SCIP-3 total scores correlated to a moderate to strong degree (r=0.63; *P*=.009), and the subtests, namely, LL and VMT, correlated to a moderate (r=0.67; *P*=.002) and strong (r=0.71; *P*=.001) degree, respectively. The usability evaluation of the ICAT system revealed high scores above 4 for system usefulness, interface quality, and overall usage. The information quality was rated lower by the English-speaking participants (3.84), compared with the Danish participants (4.11), which may indicate that the English instructions of the ICAT tests should be revised. The insignificant error rates of ASR, as calculated for the Danish and English responses (6.3% and 17.8%, respectively), indicate a promising future of ASR, particularly for Danish-speaking patients who will be the primary users of the ICAT. According to the results obtained from the recent THINC-it validity study on healthy subjects [10], the 2 cognitive games called Trails (executive function, r=0.74) and Codebreaker (attention, working memory, and executive function, r=0.63) revealed strong to moderate convergent validity, respectively, whereas Symbol Check (working memory, r=0.19) and Spotter (attention, r=0.44) showed low validity. In our study, the ICAT subtest for psychomotor speed also showed moderate concurrent validity, as did the subtest for verbal memory. However, the subtests tapping into working memory and executive skills did not correlate with the original SCIP tasks, which might be because of a suboptimal design of these tests or the small sample size. The ICAT may be an alternative to the THINC-it, which is the most recent cognitive screening tool developed specifically for UD patients. The analysis between the total scores of the SCIP and ICAT showed moderate to strong correlations (r=0.63) in contrast to the moderate concurrent validity (r=0.42) of the THINC-it composite. The higher concurrent validity and the automatic real-time verbal memory assessment via ASR are thus the advantages of the ICAT.

The lack of statistical significance between task 2 of the SCIP and ICAT might be because of the replacement of the oral countdown task with the sorting module because (1) 2 participants in Study 2 mentioned that the ICAT CR subtest was easier compared with the paper-based SCIP CR and (2) participants received high scores in the ICAT CR subtest for both studies (Figure 7), which may indicate a ceiling effect for this task. The insignificant coefficients may indicate that the participants' cognitive load in the ICAT sorting module was less than the countdown task in the SCIP CR subtest. Hence, the ICAT will require additional modifications before conducting a larger validation study of healthy individuals and patients with affective disorders.

The lack of statistical significance in the DLL task was unexpected because the ASR component was the same for both the ICAT LL and DLL subtests. When doing a poststudy analysis of the recorded data, we found that poor recognition was mainly rooted in 2 factors: (1) the subject did very fast recalls of the words and uttered them right after each other, with no or limited pauses in-between each word or (2) the subject spoke very quietly and far from the microphone. Therefore, the lack of a statistically significant correlation between the ICAT and SCIP DLL tasks might be because of the various ways in which the participants repeated the recalled words. It was previously shown that speech recognition did not perform well for non-native speakers [22], which perhaps justified the higher WER of the English responses for the participants of Study 1 (11 non-native English speakers). The analysis of the ASR of the English-speaking subjects would be more robust if we could recruit more English-speaking participants, especially native speakers. The words which received the lowest accuracy (*bed* and *hat* from the English list and *seng* and *brev* from the Danish list) should be replaced with other words provided in the SCIP manual. The lower ratio of deletion error indicated that ASR received most of the verbal responses in the ICAT LL and DLL subtests.

Digitizing validated paper-and-pencil tests requires effort in prototyping, iterative design, implementation, and evaluation. The ICAT is the first Web-based application designed based on the SCIP as a gold-standard cognitive test battery. Moreover, to our knowledge, none of the existing digital cognitive assessment tools provides a real-time assessment of verbal memory. Taking it all together, the ICAT is a novel digital tool for cognitive assessment. The feasibility of the ICAT reported in this study indicates a promising use for out-of-clinical assessment. The ultimate goal of our research is to introduce the ICAT as a brief cognitive assessment tool for remote administration and the assessment of affective disorder patients.

Implications for Future Development

On the basis of our observations, the sorting module in the ICAT CR subtest was difficult to use for most of the participants. In addition to this issue, the analysis did not show significant correlations between the SCIP and ICAT CR subtest. Consequently, the sorting module in the ICAT should be redesigned to resemble the SCIP CR task better, for example, with a speech interface, because changing the type of the interface was perhaps the primary reason for the insignificant correlation coefficient.

To mitigate the speech recognition problems, the ICAT should incorporate detailed instructions and tutorials that teach and train users how to speak loudly, clearly, and close to the microphone. Moreover, the speech recognition should be able to detect when users repeat the words too fast or quietly and then instruct them to slow down or speak more clearly. The goal is to enable the ICAT to be administered by the patient, and hence, a strong emphasis should be placed on providing self-explanatory instructions and tutorials to the users.

Limitations

This is the feasibility study of the ICAT with a limited number of participants. Despite the promising results, there are a set of limitations of the study. First, the evaluation of the ASR for English-speaking participants was limited because of the few number of native English speakers. We did not evaluate the English proficiency of the participants of Study 1 to examine whether or not the ASR recognition error was because of their English proficiency level. Second, the think-aloud method was not practical, especially during the ICAT LL and DLL subtests in which users repeated their recalled words. As cognitive tests demand mental effort, it was hard for the participants to verbalize their thoughts during the test. Hence, an implicit or objective approach for recognizing participants' interaction with the system throughout the test would be more practical. Third, the nonsignificant coefficients of the executive function and working memory according to Pearson correlation analysis might be because of the modest sample size of Study 2. Fourth, the SCIP VF task and ICAT WAIS LNS task do not translate directly into exactly the same aspect of executive functions. VF performance has been found to correlate with fluid reasoning and shifting aspects of executive function [23], whereas WAIS LNS more specifically measures working memory [24]. It is worth mentioning that currently, Google's ASR converts any arbitrary word to the closest meaningful word. Hence, the rationale for replacing the VF task with the WAIS LNS task in the ICAT was the possibility of misinterpretation caused by using the ASR technology. Finally, this pilot study included only healthy control participants. The ICAT is intended to be used for cognition screening in patients with mood disorders. On the basis of the preliminary findings from this report, our group is, therefore, in the process of validating a slightly revised version of the ICAT in patients with mood disorders.

Conclusions

The ICAT is a patient-administered, Web-based tool to screen for cognitive impairment in patients with affective disorders. The results indicate that the ICAT is a good initial step toward building a digital modified cognitive assessment tool based on the SCIP. The high values of the psychometric factors derived from the PSSUQ scores present the ICAT as a usable and useful tool. The use of real-time ASR during the immediate and delayed recall gave a WER of 17.8% and 6.3% for English and Danish responses, respectively. On the basis of the results and insights derived from this study, future optimization and further validation of the ICAT are now warranted.

Acknowledgments

PH is an early stage researcher of technology enabled mental health (TEAM) for young people. TEAM has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 722561. This study was financially supported by the CACHET, Denmark. KWM acknowledges the Lundbeck Foundation and Weimann Foundation for their contributions to her salary. The authors wish to thank all of the study participants.

Conflicts of Interest

KWM reports having received consultancy fees from Lundbeck, Allergan, and Janssen in the past 3 years. LVK has been a consultant for Lundbeck for the past 3 years. The remaining authors report no conflicts of interest.

Multimedia Appendix 1

User experience design of the (ICAT) internet-based cognitive assessment tool.

[PDF File (Adobe PDF File), 1MB-Multimedia Appendix 1]

References

1. Vieta E, Berk M, Schulze TG, Carvalho AF, Suppes T, Calabrese JR, et al. Bipolar disorders. Nat Rev Dis Primers 2018 Dec 8;4:18008. [doi: 10.1038/nrdp.2018.8] [Medline: 29516993]
2. Tourjman SV, Juster RP, Purdon S, Stip E, Kouassi E, Potvin S. The screen for cognitive impairment in psychiatry (SCIP) is associated with disease severity and cognitive complaints in major depression. Int J Psychiatry Clin Pract 2019 Mar;23(1):49-56. [doi: 10.1080/13651501.2018.1450512] [Medline: 29553848]
3. Miskowiak KW, Burdick KE, Martinez-Aran A, Bonnin CM, Bowie CR, Carvalho AF, et al. Methodological recommendations for cognition trials in bipolar disorder by the International Society for Bipolar Disorders Targeting Cognition Task Force. Bipolar Disord 2017 Dec;19(8):614-626 [FREE Full text] [doi: 10.1111/bdi.12534] [Medline: 28895274]
4. Miskowiak KW, Burdick KE, Martinez-Aran A, Bonnin CM, Bowie CR, Carvalho AF, et al. Assessing and addressing cognitive impairment in bipolar disorder: the International Society for Bipolar Disorders Targeting Cognition Task Force recommendations for clinicians. Bipolar Disord 2018 Dec;20(3):184-194. [doi: 10.1111/bdi.12595] [Medline: 29345040]
5. Jensen JH, Støttrup MM, Nayberg E, Knorr U, Ullum H, Purdon SE, et al. Optimising screening for cognitive dysfunction in bipolar disorder: validation and evaluation of objective and subjective tools. J Affect Disord 2015 Nov 15;187:10-19. [doi: 10.1016/j.jad.2015.07.039] [Medline: 26301477]
6. Ott CV, Bjertrup AJ, Jensen JH, Ullum H, Sjælland R, Purdon SE, et al. Screening for cognitive dysfunction in unipolar depression: validation and evaluation of objective and subjective tools. J Affect Disord 2016 Jan 15;190:607-615. [doi: 10.1016/j.jad.2015.10.059] [Medline: 26583350]
7. Cambridge Cognition. CANTAB Mobile. Timely detection of memory impairmentURL:http://www.cambridgecognition.com/products/digital-healthcare-technology/cantab-mobile/[WebCite Cache ID 75GimIcUD]
8. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS vital signs. Arch Clin Neuropsychol 2006 Oct;21(7):623-643 [FREE Full text] [doi: 10.1016/j.acn.2006.05.007] [Medline: 17014981]

9. Davis MT, DellaGioia N, Matuskey D, Harel B, Maruff P, Pietrzak RH, et al. Preliminary evidence concerning the pattern and magnitude of cognitive dysfunction in major depressive disorder using cogstate measures. J Affect Disord 2017 Dec 15;218:82-85 [FREE Full text] [doi: 10.1016/j.jad.2017.04.064] [Medline: 28460315]
10. Harrison JE, Barry H, Baune BT, Best MW, Bowie CR, Cha DS, et al. Stability, reliability, and validity of the THINC-it screening tool for cognitive impairment in depression: a psychometric exploration in healthy volunteers. Int J Methods Psychiatr Res 2018 Dec;27(3):e1736 [FREE Full text] [doi: 10.1002/mpr.1736] [Medline: 30088298]
11. Pakhomov SV, Marino SE, Banks S, Bernick C. Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency. Speech Commun 2015 Dec 1;75:14-26 [FREE Full text] [doi: 10.1016/j.specom.2015.09.010] [Medline: 26622073]
12. König A, Linz N, Tröger J, Wolters M, Alexandersson J, Robert P. Fully automatic speech-based analysis of the semantic verbal fluency task. Dement Geriatr Cogn Disord 2018;45(3-4):198-209. [doi: 10.1159/000487852] [Medline: 29886493]
13. Toth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Banreti Z, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr Alzheimer Res 2018;15(2):130-138 [FREE Full text] [doi: 10.2174/1567205014666171121114930] [Medline: 29165085]
14. Tröger J, Linz N, König A, Robert P, Alexandersson J. Telephone-Based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth'18; May 21-24, 2018; New York, USA p. 59-66. [doi: 10.1145/3240925.3240943]
15. Bonnín CM, Martínez-Arán A, Torrent C, Pacchiarotti I, Rosa AR, Franco C, et al. Clinical and neurocognitive predictors of functional outcome in bipolar euthymic patients: a long-term, follow-up study. J Affect Disord 2010 Feb;121(1-2):156-160. [doi: 10.1016/j.jad.2009.05.014] [Medline: 19505727]
16. Hafiz P, Miskowiak KW, Kessing LV, Bardram JE. Design and Implementation of a Web-Based Application to Assess Cognitive Impairment in Affective Disorder. In: Proceedings of the 2018 International Conference on Digital Health. 2018 Presented at: DH'18; April 23-26, 2018; Lyon, France p. 154-155. [doi: 10.1145/3194658.3194691]
17. Balsamiq.: Balsamiq Studios, LLC URL:https://balsamiq.com/[WebCite Cache ID 73MWV5f4s]
18. Estrin D, Sim I. Health care delivery. Open mhealth architecture: an engine for health care innovation. Science 2010 Nov 5;330(6005):759-760. [doi: 10.1126/science.1196187] [Medline: 21051617]
19. Cloud Speech-to-Text. URL:https://cloud.google.com/speech-to-text/ [accessed 2018-10-04] [WebCite Cache ID 72uncPqUx]
20. Jaspers MW, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
21. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int J Hum Comput Interact 1995 Jan;7(1):57-78 [FREE Full text] [doi: 10.1080/10447319509526110]
22. Wang Z, Schultz T, Waibel A. Comparison of Acoustic Model Adaptation Techniques on Non-Native Speech. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. 2003 Presented at: ICASSP'03; April 6-10, 2003; Hong Kong, China. [doi: 10.1109/ICASSP.2003.1198837]
23. Aita S, Boettcher A, Slagel B, Holcombe J, Espenan M, King M, et al. C-42the relation between verbal fluency and executive functioning: an exploratory factor analysis (EFA) approach. Arch Clin Neuropsychol 2016 Aug 31;31(6):656. [doi: 10.1093/arclin/acw043.191]
24. Crowe SF. Does the letter number sequencing task measure anything more than digit span? Assessment 2000 Jun;7(2):113-117. [doi: 10.1177/107319110000700202] [Medline: 10868248]

Abbreviations

ASR: automatic speech recognition
BD: bipolar disorder
CACHET: Copenhagen Center for Health Technology
CARP: CACHET Research Platform
CNSVS: central nervous system vital sign
CR: consonant repetition
DLL: delayed list learning
GDPR: general data protection regulation
ICAT: internet-based cognitive assessment tool
ISBD: International Society for Bipolar Disorder
LL: list learning
MCI: mild cognitive impairment
mHealth: mobile health
PSSUQ: poststudy system usability questionnaire
SCIP: screen for cognitive impairment in psychiatry
TEAM: technology enabled mental health

UD: unipolar disorder
UI: user interface
UX: user experience
VF: verbal fluency
VMT: visuomotor tracking
WAIS LNS: Wechsler Adult Intelligence Scale letter-number sequencing
WER: word error rate

Edited by G Eysenbach; submitted 04.03.19; peer-reviewed by L Germine, O Ajilore, J Zulueta; comments to author 06.04.19; revised version received 30.05.19; accepted 29.06.19; published 26.07.19

Please cite as:
Hafiz P, Miskowiak KW, Kessing LV, Elleby Jespersen A, Obenhausen K, Gulyas L, Zukowska K, Bardram JE
The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study
JMIR Form Res 2019;3(3):e13898
URL: http://formative.jmir.org/2019/3/e13898/
doi: 10.2196/13898
PMID: 31350840

CHAPTER 10

Papers Related to UbiCAT

10.1 Design and Formative Evaluation of Cognitive Assessment Apps for Wearable Technologies

Authors: **Pegah Hafiz** and Jakob E Bardram

Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers

DOI: https://doi.org/10.1145/3341162.3347077

Design and Formative Evaluation of Cognitive Assessment Apps for Wearable Technologies

Pegah Hafiz & Jakob E. Bardram
Department of Health Technology, Technical University of Denmark
Lyngby, Denmark
[pegh,jakba]@dtu.dk

ABSTRACT

Cognitive functioning is a crucial aspect of the individual's mental health and it affects human's daily activities. We have developed the Ubiquitous Cognitive Assessment Tool (UbiCAT) including three cognitive assessment apps on the Fitbit smartwatch. In this paper, we present the design and formative evaluation of the UbiCAT apps conducted with 5 participants who had a background in design and/or human-computer interaction. Moreover, we investigated the adoption of the wearable devices by our participants.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile devices*;

KEYWORDS

cognition; cognitive function; wearable technology; application; stroop test

ACM Reference Format:
Pegah Hafiz & Jakob E. Bardram. 2019. Design and Formative Evaluation of Cognitive Assessment Apps for Wearable Technologies. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341162.3347077

1 INTRODUCTION

Cognition is a core function to the human daily activities. Cognitive measures include working memory, verbal memory, attention, psychomotor speed and executive function. Human cognition may fluctuate during the day. For example, an individual may have a better attention in the morning. The cognitive fluctuation level may vary among individuals depending on several factors including their age and mental workload. Hence, a personalized model of cognition seems practical for the individuals to reflect on their cognitive performance during the day.

Human cognition and alertness have been previously investigated via mobile apps. Dingler et al. [4] assessed alertness using three short tasks, namely Psychomotor Vigilance Task (PVT), Go/No-Go (GNG), and Multiple Object Tracking (MOT). Abdullah et al. [1] conducted a study with 20 participants to analyze the fluctuation in individuals' alertness using PVT mobile app called PVT-Touch. Their findings showed that alertness can vary by 30% depending on the timing of the day. The standard PVT test takes 10 minutes [3] but both studies used a short version of the PVT to prevent participants' fatigue. The Cognition Kit app (https://www.cognitionkit.com/) is designed for the Apple Watch to assess mood and cognition in clinical context. The cognitive test used in this app is the 'N-back' test [6].

The Ubiquitous Cognitive Assessment Tool (UbiCAT) includes three cognitive assessment apps as well as mobile sensor data collection using Fitbit API. The UbiCAT cognitive assessment apps are short tests to conduct a research and assess cognition in the wild. Through UbiCAT project, we will collect two types of data: (i) cognitive performance of the individuals which are assessed via the apps and (ii) mobile sensor data including physical activity, heart rate, sleep, and mobility data. The ultimate goal of the UbiCAT project is to find correlation between individuals' cognitive performance and objective mobile data. The objectives of this paper are two-folded. First, the design and formative evaluation results of the UbiCAT apps will be presented. Second, the challenges regarding the wearable technologies for cognitive assessment will be discussed.

2 DESIGN AND STUDY

The UbiCAT apps are designed in a user-centered design process including 3 expert researchers who hold a PhD in computer science, experimental psychology and cognitive science. The design of the UbiCAT apps was revised after several meetings with the experts. In this section, we will first introduce the UbiCAT apps. Then, the formative evaluation of the apps will be explained in detail.

2.1 Overview of the Apps

The UbiCAT includes three apps namely Color Test (CT), Letter Test (LT) and Arrow Test (AT). Each app provides a set of short instructions as well as the test itself. All tests are timed and users should respond as fast as possible. The apps are implemented on the FitBit Ionic smartwatch. The CT is adapted from the Stroop color-word test [5], which presents a set of color names one by one, each with either the same (congruent) or different (incongruent) ink color. Figure 1 shows a screenshot of the CT with an incongruent stimuli. Users are presented with four colors to select the correct ink color of the stimuli. The Stroop test examines sustained attention and the performance measures are the Stroop effect and average response time.

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09…$15.00
https://doi.org/10.1145/3341162.3347077

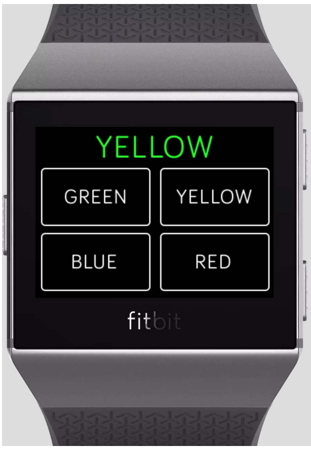


Figure 1: A screenshot of the Color test adapted from the Stroop color-word test

The LT is adapted from the 'N-back' test, where the stimulus are a set of letters shown one by one and the user is asked to memorize *N* letter back in the sequence. 'N-back' is used by psychologists and psychiatrists to assess working memory. Figure 2 shows a screenshot of the LT where N is 1 and users should tap on either 'Yes' if the current letter ('T') is the same as one letter showed back in the sequence or 'No' otherwise. This test becomes more difficult as N increases.

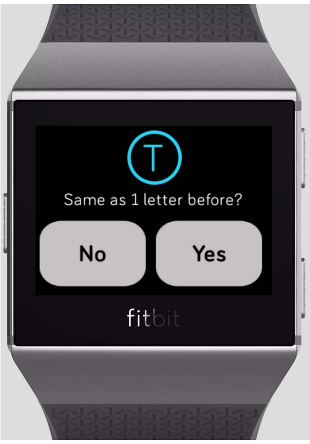

Same as 1 letter before?

Figure 2: A screenshot of the Letter test adapted from the N-back test

Choice reaction test has variations such as the GNG and stop signal test. A computer-based four-choice reaction time test is developed and tested on adults by Deary et al. [2]. The AT is a two-choice reaction time test. In this test, each stimulus is an arrow pointing either to the left or right side. A screenshot of a sample AT test is shown in Figure 3. The users are required to tap on the right/left app button in case the arrow points to the right/left.

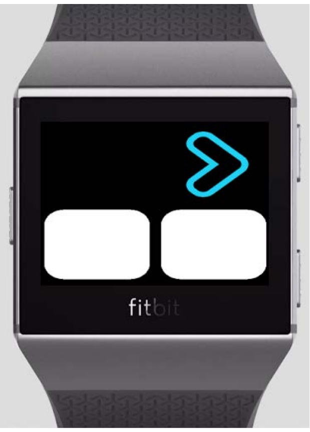


Figure 3: A screenshot of the Arrow test adapted from the two-choice reaction time test

2.2 Formative Evaluation

The aim of this study was to (i) evaluate app design and improve the user interface of the apps and (ii) explore the adoption of wearable technologies. The participants and the procedure of this study are presented below.

2.2.1 Participants. The participants of this study were selected from the researchers who had a background in app design and/or human-computer interaction at Technical University of Denmark.

2.2.2 Procedure. A consent form was handed to the participants before beginning the study. Then, we asked the participants about their experience of using a wearable device, the duration, and reason to stop using it (if any). The procedure to work with each app involved three steps. First, the participant was asked to wear the FitBit smartwatch and launch the app, read the test instructions, take a test, and view his/her score. We asked the participants to verbalize their thought as in 'think-aloud' method. Then, participants were asked to fill in a questionnaire form including 7 questions. We selected relevant questions of a reliable and objective tool called Mobile App Rating Scale (MARS) [7] to evaluate UbiCAT apps. The chosen questions were taken from three sections of the MARS namely aesthetics (3 questions), functionality (2 questions) and information (2 questions). The 5-point rating scale was used to give a score to the participants' response to each question. Finally, a semi-structured interview was conducted with each participant. The participants' interaction with the apps and the interviews were recorded.

Table 1: Usability results of the UbiCAT apps

	Aesthetics	Functionality	Information
Arrow Test	3.93 ± .61	4.6	4.4
Letter Test	4 ± .2	3.3 ± 1.84	2.6
Color Test	4 ± .2	4.2 ± .28	3.9 ± .14

3 RESULTS

Five people participated in this study (1 female, 4 male; age=28 ± 4.35), 1 with a PhD degree, 3 PhD students and 1 Master's degree student. The experiment took approximately 1 hour per participant. The interviews with the participants were transcribed after the experiment. The video recordings were also checked multiple times to identify where participants struggled with the apps.

3.1 Usability

Table 1 shows the usability results for the UbiCAT apps reported separately in terms of aesthetics, functionality and information. The range of scores in Table 1 is from 1 (lowest) to 5 (highest).

3.2 Wearable Technology Adoption

We investigated whether our participants use a wearable device or not. Of the 5 participants, 3 of them did not need to use a wearable activity tracker:

"I feel a bit overwhelmed when I have too much data about myself. I prefer not to have numbers to define myself. I do not even use a normal watch" (P1).

"[I have used] only normal watches. I have no need for it. No need to track the data" (P2).

"I never felt I have to. I already tracked the things that I wanted on my smartphone on the street or park and I do not think I need it" (P3).

Two of the participants had already used at least one wearable device for various time periods: *"It (Apple watch) did not really give me anything. It was too obtrusive and pervasive. You are always reminded of something. A classic mechanical watch does not disturb me since it does not collect data and it is passive"* (P4 used Basic Pick and Apple Watch for several months).

"I used Nokia smartwatch for six months but I felt I was motivated without it" (P5 used Nokia steel HR).

3.3 Feedback

UbiCAT apps give feedback to the participants when a user responds to a stimuli. A sample screenshot of the feedback to a correct response in the CT test is shown in Figure 4. We received the following comment regarding the feedback to the user's response:

"I am not sure how feedback affects my performance. How could it affect? If it indicates that I entered a response...for couple of times it distracted me to think if I was right or wrong. It is not about the amount of the feedback. Cognitively, I might be doing something else. Maybe I want to reduce my error rate instead of responding as fast as possible" (P4).

The rest of the participants did not report any issue in regard to the app feedbacks during the tests.

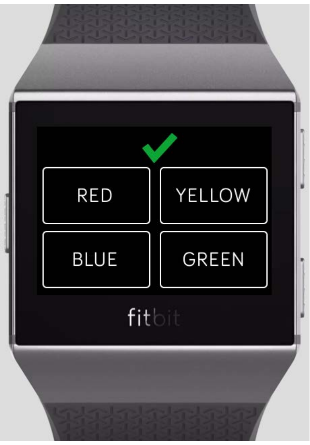


Figure 4: A screenshot of a feedback to a correct response in the Color test app

3.4 Test Score

All UbiCAT apps use the same user interface to display a score after the test. Upon finishing a test, the number of correct responses is displayed on the smartwatch screen. During the user-centered design meetings, we decided to make it as simple as possible due to the negative impact of the low score on the individual's mood. The following question was asked during the interviews: *"Did your score at the end of the test help you to understand your performance? If yes, how?"*

All of the participants mentioned that a maximum score is essential for them to understand their performance, for instance, 24 (number of correct responses) out of 30 (maximum score). We perceived that quantifying user's performance is essential.

4 CONCLUSION AND FUTURE WORK

The usability study showed promising outcomes in terms of aesthetics, functionality and information of the UbiAT apps. The score of the information section for the LT gave average score which indicates that the instruction set of this test should be improved. Through this study, we assessed the usability of the UbiCAT apps and investigated the participants' interaction with the wearable devices. The participants' comments regarding wearable devices will help us in improving the usability of the wearable devices in particular smartwatches. We have planned to conduct a study to evaluate the UbiCAT apps with more participants.

ACKNOWLEDGMENTS

This project is developed as a part of PH's PhD. PH is an early stage researcher of Technology Enabled Mental Health (TEAM) for Young People. TEAM has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 722561. We wish to thank our participants.

REFERENCES

[1] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 178–189.

[2] Ian J Deary, David Liewald, and Jack Nissan. 2011. A free, easy-to-use, computer-based simple and four-choice reaction time programme: the Deary-Liewald reaction time task. *Behavior research methods* 43, 1 (2011), 258–268.

[3] David F Dinges and John W Powell. 1985. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior research methods, instruments, & computers* 17, 6 (1985), 652–655.

[4] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building Cognition-Aware Systems: A Mobile Toolkit for Extracting Time-of-Day Fluctuations of Cognitive Performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 47.

[5] Charles J Golden and Shawna M Freshwater. 1978. Stroop color and word test. (1978).

[6] Wayne K Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology* 55, 4 (1958), 352.

[7] Stoyan R Stoyanov, Leanne Hides, David J Kavanagh, Oksana Zelenko, Dian Tjondronegoro, and Madhavan Mani. 2015. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth* 3, 1 (2015), e27.

10.2 The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study

Authors: **Pegah Hafiz** and Jakob E Bardram

Journal of Medical Internet Research [**In press**]

DOI: 10.2196/17506

Original Paper

The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study

Pegah Hafiz[1,2], MSc; Jakob Eyvind Bardram[1,2], MSc, PhD

[1]Digital Health Section, Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark

[2]Copenhagen Center for Health Technology, Kongens Lyngby, Denmark

Corresponding Author:
Pegah Hafiz, MSc
Digital Health Section
Department of Health Technology
Technical University of Denmark
Richard Petersens Plads
Building 324, 2nd Floor, Room 270
Kongens Lyngby, 2800
Denmark
Phone: 45 91858371
Email: pegh@dtu.dk

Abstract

Background: Cognitive functioning plays a significant role in individuals' mental health, since fluctuations in memory, attention, and executive functions influence their daily task performance. Existing digital cognitive assessment tools cannot be administered in the wild and their test sets are not brief enough to capture frequent fluctuations throughout the day. The ubiquitous availability of mobile and wearable devices may allow their incorporation into a suitable platform for real-world cognitive assessment.

Objective: The aims of this study were threefold: (1) to evaluate a smartwatch-based tool for the assessment of cognitive performance, (2) to investigate the usability of this tool, and (3) to understand participants' perceptions regarding the application of a smartwatch in cognitive assessment.

Methods: We built the Ubiquitous Cognitive Assessment Tool (UbiCAT) on a smartwatch-based platform. UbiCAT implements three cognitive tests—an Arrow test, a Letter test, and a Color test—adapted from the two-choice reaction-time, N-back, and Stroop tests, respectively. These tests were designed together with domain experts. We evaluated the UbiCAT test measures against standard computer-based tests with 21 healthy adults by applying statistical analyses significant at the 95% level. Usability testing for each UbiCAT app was performed using the Mobile App Rating Scale (MARS) questionnaire. The NASA-TLX (Task Load Index) questionnaire was used to measure cognitive workload during the N-back test. Participants rated perceived discomfort of wearing a smartwatch during the tests using a 7-point Likert scale. Upon finishing the experiment, an interview was conducted with each participant. The interviews were transcribed and semantic analysis was performed to group the findings.

Results: Pearson correlation analysis between the total correct responses obtained from the UbiCAT and the computer-based tests revealed a significant strong correlation (r=.78, P<.001). One-way analysis of variance (ANOVA) showed a significant effect of the N-back difficulty level on the participants' performance measures. The study also demonstrated usability ratings above 4 out of 5 in terms of aesthetics, functionality, and information. Low discomfort (<3 out of 7) was reported by our participants after using the UbiCAT. Seven themes were extracted from the transcripts of the interviews conducted with our participants.

Conclusions: UbiCAT is a smartwatch-based tool that assesses three key cognitive domains. Usability ratings showed that participants were engaged with the UbiCAT tests and did not feel any discomfort. The majority of the participants were interested in using the UbiCAT, although some preferred computer-based tests, which might be due to the widespread use of personal computers. The UbiCAT can be administered in the wild with mentally ill patients to assess their attention, working memory, and executive function.

(JMIR Mhealth Uhealth 2020;8(6):e17506) doi: 10.2196/17506

KEYWORDS

cognition; memory; response time; attention; Stroop task; wearable devices; mobile phone

Introduction

Background

Wearable devices provide an opportunity for users to collect their personal data. A recent empirical study determined that fashnology, individuals' attitudes, and risk context were the most influential factors in adoption of wearable devices for quantified self-tracking purposes [1]. Wrist-worn devices, particularly smartwatches, are becoming more popular. Usefulness and visibility are the two major reasons that people adopt a smartwatch [2]. Smartwatches are lightweight and portable, which makes them easy for people to wear and use in almost every context, while some people may not carry their smartphones when they go for a walk or run. Moreover, platforms, including Fitbit OS (operating system), Apple Watch's watchOS, and Google's Wear OS, support building stand-alone apps that run without connecting to a smartphone. The application programming interface (API) of some smartwatches allow sensor data collection in the wild, including physiological and behavioral data, such as sleep, heart rate variability, mobility, and location. King and Saffarzadeh reviewed the application of smartwatches in 27 health-related studies [3]. Their findings show that activity monitoring, chronic disease self-management, nursing or home-based care, and health care education are the current smartwatch-based applications in health care. Hence, smartwatches are suitable devices to assist researchers in developing stand-alone health care–related apps, as well as for collecting sensor data in the wild.

Cognitive functioning is a crucial aspect of mental health and determines the quality of individuals' daily activities. According to Lyon et al, impairment in attention, memory, and executive function may cause problems at school or work [4]. Moreover, previous studies have shown daily fluctuations in alertness [5], working memory [6], and executive skills [7]. Quantifying cognitive performance may help individuals reflect on their own fluctuations. For instance, students can track their alertness levels to select appropriate times of day to schedule their attention-demanding tasks. Besides healthy individuals, mentally ill patients also suffer from cognitive dysfunction, such as dementia [8], bipolar disorder [9,10], attention deficit hyperactivity disorder (ADHD) [11], and schizophrenia [12]. Monitoring cognitive performance can thus help patients in scheduling their follow-up visits in case of significant degradation in their cognitive functioning, as it may indicate the onset of their illness.

Digital cognitive screening tools have been designed for different technological platforms, targeting both mentally ill patients and healthy individuals. The Cambridge Neuropsychological Test Automated Battery (CANTAB) Mobile [13], the Internet-based Cognitive Assessment Tool (ICAT) [14], the THINC-integrated tool (THINC-it) [15], MyCognition Quotient (MyCQ) [16], CogState [17], and the Brief Assessment of Cognition in Schizophrenia (BACS) [18] are some examples of the validated cognitive test batteries administered on a computer or tablet. The existing cognitive test batteries are administered at a certain time in a controlled condition. Such cognitive tools are not feasible for long-term frequent monitoring and assessment of cognitive functioning, since (1) it takes at least 15 minutes to complete a set of tests and (2) the tests are taken in a controlled condition without any distraction, for example, a silent room. However, according to previous studies [19,20], it is crucial to assess cognitive functioning in real-life settings for frequent and continuous monitoring of the individuals.

Ecological momentary assessment (EMA) [21] and the experience-sampling method (ESM) [22] were developed to overcome the bias in delivering retrospective self-reports by study participants. Both methodologies provide an opportunity to collect psychological and clinical measures of behavior, cognition, and emotion in situ [23]. Unobtrusive cognitive tests instead of subjective ratings may improve the accuracy of EMA and the ESM in longitudinal studies.

Taking together, a ubiquitous tool providing continuous and frequent assessment of the individuals' in-the-wild cognitive performance would be an important approach for real-world psychometric research and diagnosis.

Previous Studies

Overview

The application of neuropsychological tests on mobile platforms previously showed promising outcomes [19,24]. In this section, an overview of the previous studies on digital cognitive tests developed for smartphones and smartwatches is presented. Commercial cognitive training mobile apps with no evidence of validity were excluded.

Smartphone-Based Tools

A research platform called iVitality includes a smartphone app with five cognitive tests, namely Memory-Word, Trail Making, Stroop, Reaction Time, and N-back. Jongstra et al conducted a study with 151 healthy individuals to examine feasibility and validity of the iVitality platform over 6 months [25]. According to the results of their validation study, the Stroop and Trail Making tests correlated moderately (r=.5 and r=.4, respectively) with the conventional tests. The authors did not validate the rest of the cognitive tests against their corresponding baseline measures, due to the difference between the raw scores of the smartphone tests and conventional tests. The Color-Shape Test (CST) is a smartphone-based app designed to measure cognitive processing speed and attention in the elderly population. The validity of the CST was examined against the Uniform Data Set (UDS) neuropsychological test battery in an experiment by Brouillette et al with 57 individuals who did not have dementia [26]. Their findings showed a significant correlation between CST scores and global cognition with the Mini-Mental State Examination (r=.52), Digit Span (r=.43), the Trail Making test (r=-.65), and the Digit Symbol test (r=.51). However, the CST scores did not correlate with verbal fluency tasks. Tieges et al conducted a study with 20 delirium patients to assess the feasibility of a smartphone-based app called the DelApp against a computerized device called the Edinburgh Delirium Test Box (EDTB) [27]. The authors found no significant difference between the scores of the DelApp and the EDTB (P=.41). Pal et al used a mobile app called the Neurophone, which includes

N-back, Stop Signal, and Stroop tests, to evaluate the cognitive performance of 20 healthy and 16 methamphetamine users against a validated computerized tool [28]. The Stop Signal test results could not be compared to the computerized tests due to the different parameters used by phone- and computer-based tests, while the scores of the N-back test on both platforms were similar. The authors used speech recognition in the Stroop test of their mobile app to detect the correct response time (RT). However, due to the inaccuracy of the speech recognition, the test results of the computer- and phone-based tests were not comparable. Dingler et al developed a smartphone-based tool including three short cognitive tasks, namely the psychomotor vigilance task (PVT), the go/no-go task, and the multiple object tracking task [29]. The authors conducted an in-the-wild study to assess the alertness of 12 participants over 9 days, on average. Although the short version of the PVT was validated before by Basner et al [30], the go/no-go and the multiple object tracking tasks were not tested against a computer- or paper-based neuropsychological test.

Smartwatch-Based Tool

Cormack et al developed a tool called the Cognition Kit for the Apple Watch, including a variation of the N-back test adapted from CANTAB's N-back along with self-reports of mood using a short questionnaire [31]. The authors conducted feasibility and validation studies of the Cognition Kit with 30 depressed patients. According to their validation study results, N-back test performance correlated with CANTAB’s rapid visual information processing task (P≤.01, r=.5).

Gaps in the Literature

The study conducted by Dingler et al [29] was the only work that introduced a smartphone-based toolkit for doing research on in situ alertness. The rest of the smartphone-based apps were developed to deliver personal cognitive assessment tools without collecting mobile data. So far, the Cognition Kit is the only smartwatch-based tool exclusively assessing working memory through the N-back test. A limited number of cognitive measures provided by mobile tools, as well as a lack of studies in exploring the potential of smartwatches in measuring in-the-wild cognition, led us to build the Ubiquitous Cognitive Assessment Tool (UbiCAT). Our tool has three smartwatch-based cognitive tests measuring three key cognitive domains, namely attention, working memory, and executive function. The UbiCAT tests, along with smartwatch-based sensor data collection, allow researchers to analyze associations between individuals’ cognitive, physiological, and behavioral features toward identifying digital biomarkers of human cognitive functioning and conducting psychometric research in the wild.

Goals of This Study

Through this study, we will (1) evaluate the cognitive measures of the UbiCAT apps against state-of-the-art computer-based tools, (2) assess the usability of the UbiCAT tests, and (3) understand participants' perceptions about smartwatch apps for assessing cognition.

Methods

In this section, we first provide details of the design and functionality of the UbiCAT apps; we then explain the study in detail.

Design Methods

Overview

The UbiCAT includes three smartwatch-based apps; each is a cognitive test that measures a certain cognitive domain. We considered three inclusion criteria for the UbiCAT tests: (1) the tests should measure memory, attention, and executive function, since fluctuations in these domains may negatively affect individuals’ work or study performance, (2) each test should be able to be adapted for the limited screen size of the smartwatch, and (3) each test should not require a microphone or speaker, which are essential in verbal recall tests. Taking these together, we selected a two-choice reaction-time test [32] to measure attention, the Stroop color-word test [33] to measure attention and executive function, and the N-back test [34] to examine working memory. The three tests contribute to short assessments, as it takes approximately 5 minutes to take the UbiCAT tests.

Three experts who each hold a doctoral degree within cognitive psychology and human-computer interaction were involved in the design process. First, the initial design of the aforementioned tests was sketched on paper. Based on detailed analysis of the available smartwatch hardware platforms, the Fitbit Ionic device was selected. Second, functional prototypes for each test were implemented separately and tested on the smartwatch. Individuals with different finger sizes were asked to work with the apps to adjust the size of the app buttons and text. The Fitbit design guidelines were also considered during the prototyping phase. The components of the UbiCAT apps were revised several times after meetings with the domain experts. Overall, the design and implementation process took 4 months. Third, a formative evaluation study of the earlier versions of the UbiCAT apps was conducted with 5 participants aimed to examine the usability of the apps and understand participants' adoption of wrist-worn devices [35]. The findings of the formative evaluation study helped us improve the user interface and functionality of the apps.

The UbiCAT Cognitive Tests

Overview

Three stand-alone apps were built for the Fitbit smartwatch. Each test takes less than 2 minutes to complete. We selected the following names for the UbiCAT apps to simplify memorizing the apps for the users: *Arrow test* (two-choice reaction-time test), *Letter test* (N-back test), and *Color test* (Stroop color-word test). An outline of the UbiCAT apps is presented in the following sections and snapshots are shown in Figures 1-3.

Arrow Test

The Arrow test presents a sequence of rightward or leftward arrows to the user one by one. The user is required to select the correct direction of each arrow by tapping on either the left or right app button. The position of each arrow can be on the left or right side of the screen. Figure 1 shows a snapshot of the Arrow test where the correct response to this stimulus is the app button on the right side.

Letter Test

In the Letter test, a sequence of English alphabet letters are displayed to the user. Depending on the value of N, the user is supposed to determine whether the current stimulus is the same as the N letter, or N letters, back in the sequence or not. The value of N determines the difficulty level and is unchanged during an entire trial. Figure 2 shows a snapshot of the 2-back test, where N is equal to 2.

Color Test

The names of four colors, for example *RED*, with either the same or different ink color are the stimuli of the Color test. A congruent stimulus has the same color as its meaning, while an incongruent stimulus has a different color. The task of the user is to select the ink color of each stimulus by tapping on the app button labeled with the color name. Figure 3 presents an incongruent stimulus. Here, the correct response is the *GREEN* app button in the bottom-left corner.

Figure 1. A sample test taken from the UbiCAT (Ubiquitous Cognitive Assessment Tool) Arrow test. The stimuli is the rightward arrow and the app buttons on both sides capture the direction of the arrow.

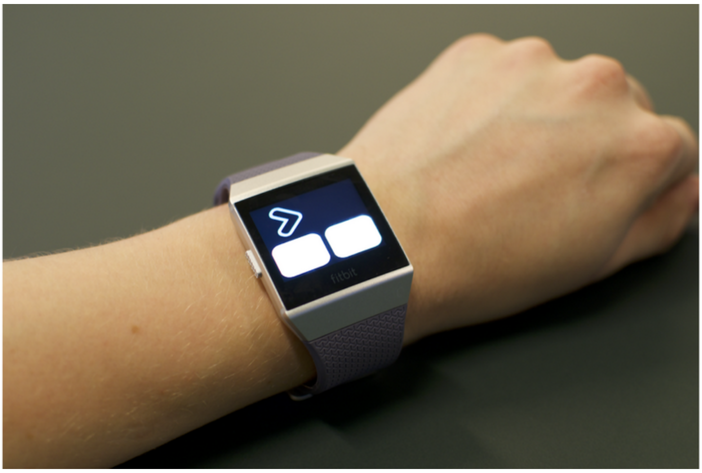


Figure 2. A sample test taken from the UbiCAT (Ubiquitous Cognitive Assessment Tool) Letter test, 2-back task. The participant should indicate whether “T” appeared 2 letters back in the sequence or not.

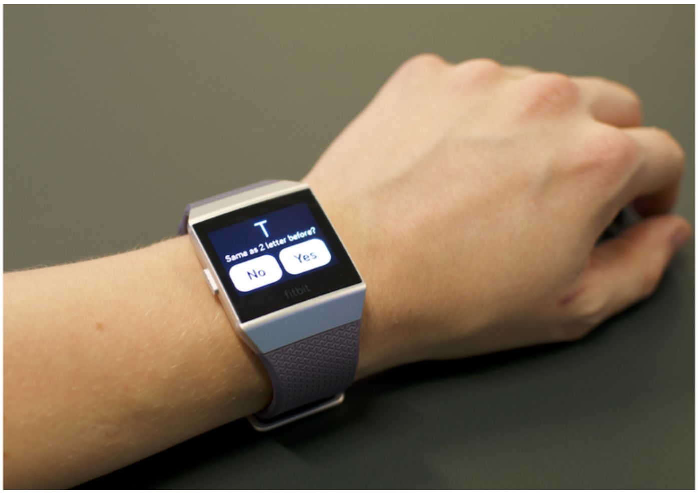
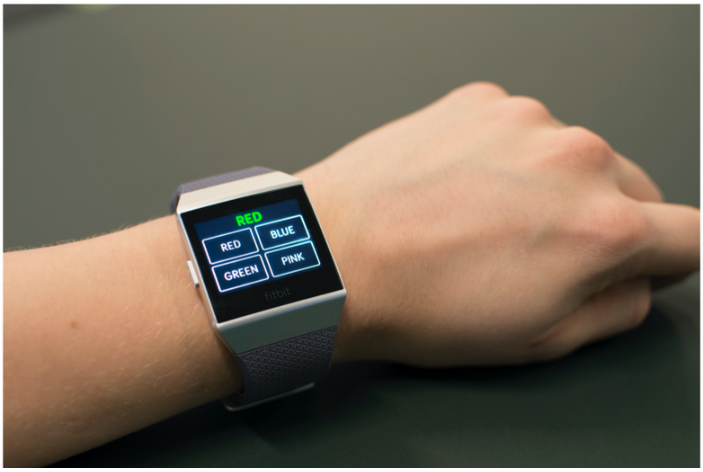


Figure 3. A sample test taken from the UbiCAT (Ubiquitous Cognitive Assessment Tool) Color test, displaying an incongruent stimuli.



Technical Specifications and Apparatus

Two validated computer-based tools, PsyToolkit [36,37] and the THINC-it application [15], were run on a MacBook Pro (15-inch Retina display, Apple Inc) during the study. A Fitbit Ionic smartwatch (1.42-inch screen, 348 × 250 pixel resolution) was used to run the UbiCAT apps. The figures in this paper were created in RStudio using the ggplot2 package [38].

Ethical Approval

The study protocol and system description were sent for approval by the Danish Ethical Committee. The study was classified as a nonclinical survey study and, hence, exempted for ethical approval (Journal-nr.: H-19086232).

Participant Recruitment

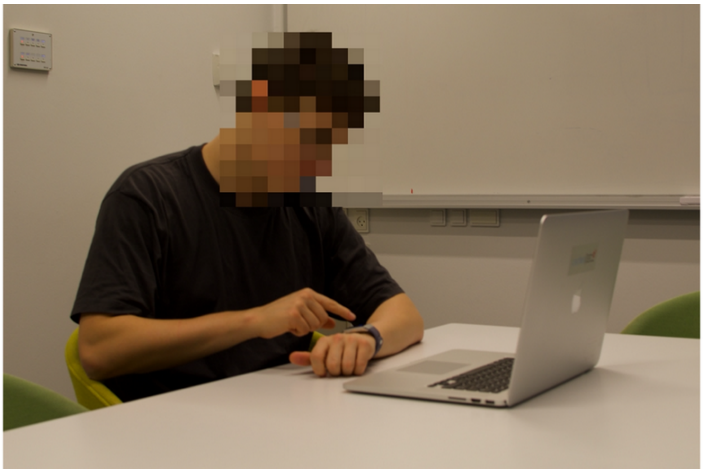
We recruited 21 healthy adults who lived in Copenhagen, Denmark, using a snowball sampling method [39]. All participants had sufficient English-language skills to read the test instructions. Participants were not eligible if they had a history of mental illness, were aged over 50 years, or had color blindness.

Procedure

Overview

All of the test sessions were performed in a silent room at the Technical University of Denmark. The study session lasted 60-75 minutes per participant. Participants were compensated with a gift card worth an amount equal to US $15 that was given at the end of the study. Prior to an experiment, the study leader (PH) informed the participant to ask for a short break between the testing sessions if needed. We measured each participant's perceived wrist discomfort after completing each of the UbiCAT tests using a 7-point Likert scale. Figure 4 shows a participant completing a UbiCAT test on the Fitbit smartwatch. A detailed description of the experiment is presented below.

Figure 4. A study participant completing a UbiCAT (Ubiquitous Cognitive Assessment Tool) test via a Fitbit Ionic smartwatch. The laptop was used to administer computer-based tests.



First, a general description of the study was given to the participant. Second, a consent form was handed to the participant. Upon signing the consent form, background information from the participant was collected, including age, gender, educational background, and preference in terms of watch-wearing wrist (ie, dominant or nondominant hand). Third, the participant was asked to perform the three UbiCAT tests one by one. Each test was administered against its corresponding computer-based test. PH explained the instructions of the computer-based tests to the participant and repeated if needed. The participant was able to read the instructions of the smartwatch-based tests in the UbiCAT by themself. The feedback displayed to the participant was the fraction of correct responses to the total responses in each UbiCAT test. The interaction of each participant with the smartwatch was video-recorded during the experiment. The order of test administration on the smartwatch and computer was counterbalanced between participants.

Previous work mentioned that some of the cognitive test results obtained from paper-based and computer-based tools could not be compared to their corresponding smartphone-based tests, due to the difference between parameters. Therefore, we chose the PsyToolkit for the Stroop color-word test and the N-back test. This tool allows researchers to program their experiments and adapt the parameters to their needs. We matched the difficulty levels of the N-back tests by changing the N and selecting the same ratio of congruent stimuli to the incongruent stimuli (1:3) in the Stroop tests. The details of our study are presented below.

Arrow Test Versus Spotter Test

All participants took the Arrow test and THINC-it Spotter test twice. The stimuli of each test on both the smartwatch and the computer was a set of 40 arrows. Each arrow was displayed on the watch for a maximum of 2000 ms. The interstimulus interval was randomly selected to be between 1000 and 3000 ms. The input of the THINC-it Spotter test was received by pressing the left or right arrow key, while the input was captured by tapping the left or right app button in the UbiCAT Arrow test. The performance measure calculated for both tests was the number of correct responses and fastest RTs.

Letter Test Versus PsyToolkit N-Back Test

The N-back test was administered separately with three difficulty levels, starting from N=1. The tests with the same difficulty level were tested against each other. For instance, 1-back in the Letter test was examined against the PsyToolkit 1-back test. The stimuli of each test was a sequence of 40 English alphabet letters displayed one by one. The time limit for the participant to respond to a stimulus was 2500 ms. Two keys were used to respond during the PsyToolkit test: “m” for yes and “n” for no. The inputs were captured on the UbiCAT Letter test by tapping on the app buttons labeled as “Yes” and “No.” The performance measures were the number of correct responses and mean RTs to the stimuli.

Color Test Versus PsyToolkit Stroop Test

All participants took each test twice. The stimulus of each test was 30 color names consisting of 7 congruent and 23 incongruent color names. The time limit was 2500 ms. Participants were required to press “b” for blue, “g” for green, “r” for red, and “y” for yellow in the PsyToolkit Stroop test. Responses were captured in the UbiCAT Color test by tapping on the app buttons labeled with the color names (see Figure 3). The pink color replaced yellow on the Fitbit smartwatch for some participants who found yellow difficult to distinguish. The performance measures of the Stroop tests were the mean RTs to the congruent and incongruent stimuli.

Usability Testing

The usability of the UbiCAT apps was assessed using the Mobile App Rating Scale (MARS) questionnaire [40]. Relevant questions concerning aesthetics, functionality, and information were selected from the MARS questionnaire (see Multimedia Appendix 1). The rating scale for each of the MARS questions ranged from 1 (the lowest score) to 5 (the highest score).

Perceived Cognitive Workload

Each N-back task was preceded by the NASA-TLX (Task Load Index) questionnaire [41] to quantify participants' perceived cognitive workloads using a 7-point Likert scale. The following subscales of the NASA-TLX were used: mental demand, temporal demand, performance, effort, and frustration level. It should be noted that the physical demand subscale was excluded as it was deemed irrelevant.

Follow-Up Interview

Upon finishing each experiment, a short interview was performed with each participant to investigate their subjective perception about the experiment and the UbiCAT tests, as well as their suggestions to improve the apps and/or instructions. The interviews were audio-recorded and transcribed for semantic analysis and grouping of the findings across participants.

Statistical Analysis

The Pearson correlation test was performed on the number of correct responses and mean RTs of the cognitive tests on both platforms. The paired-sample *t* test was applied on the performance measures to compare the numbers obtained from the smartwatch- and computer-based tests. One-way analysis of variance (ANOVA) was used to analyze the effect of difficulty level on the participants' test performances during the N-back test. The CI of the statistical tests was 95%. The statistical analysis was performed in JASP, version 0.11.1 (The JASP Team).

Results

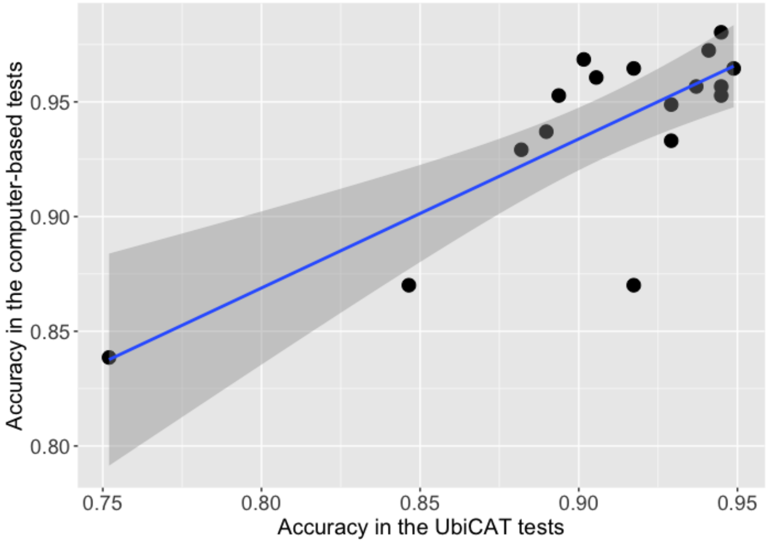
Participant Statistics

Participants were aged between 19 and 44 years (mean 26, SD 6), and 9 out of 21 participants (43%) were female. On average, participants spent 5.7 years studying at a higher-education level. Participants had diverse occupational backgrounds, including design, computer science, water engineering, construction, health care, energy, and food engineering. Of the 21 participants, 10 (48%) of them had used at least one wrist-worn device before. All participants except for 1 (20/21, 95%) wore the smartwatch on their nondominant hand.

Overall Analysis

Pearson correlation analysis revealed a significant strong correlation between the total number of correct responses obtained from the cognitive tests on the UbiCAT and computer-based tools (*r*=.78, *P*<.001). It should be noted that the scores of 4 participants of the PsyToolkit Stroop test were lost; thus, the correlation analysis between the total scores was performed for 17 participants. Figure 5 shows the total participant accuracy obtained from the UbiCAT apps versus the computer-based tools, along with the regression line. The single data point located on the bottom-left corner of Figure 5 might indicate an outlier; however, we did not remove this sample point, since it is normal that the abilities of the individuals are different from each other.

Figure 5. Overall participant accuracy in the three cognitive tests. Each black dot represents results from one participant. The blue line is the regression line and the shaded region is the CI. UbiCAT: Ubiquitous Cognitive Assessment Tool.



Two-Choice Reaction-Time Tests

The Pearson correlation analysis that was applied on the average of correct responses in the two trials of the Arrow test and Spotter test and the participants' fastest RTs on both platforms is presented in Table 1. Figure 6 shows the box plots of the number of correct responses for both platforms during each trial. Figure 7 shows the box plots of the participants' fastest RTs calculated for both trials of the two-choice reaction-time tests. We applied the paired-sample Student *t* test and it revealed that the fastest RTs obtained from the Arrow test in both trials were not statistically different (t_{20}=-1.266, *P*=.22). The average of the participants' fastest RTs in the Arrow test were statistically higher than in the Spotter test (t_{20}=10.84, *P*<.001).

N-Back Test

Figures 8 and 9 show the number of correct responses and the mean RTs of the participants, respectively, during the 1-back, 2-back, and 3-back tests in the Letter test and the PsyToolkit N-back test. Pearson correlation analysis was performed on the number of correct responses and the mean RTs for each difficulty level between the Letter test and PsyToolkit N-back test. The results are presented in Table 2.

Table 1. Correlation analysis between performance measures in the Arrow test and the Spotter test.

Performance measure	Pearson *r*	*P* value
Average of correct responses	.61	.003
Fastest response times	.24	.30

Figure 6. Box plots of participants' number of correct responses during the two-choice reaction-time tests.

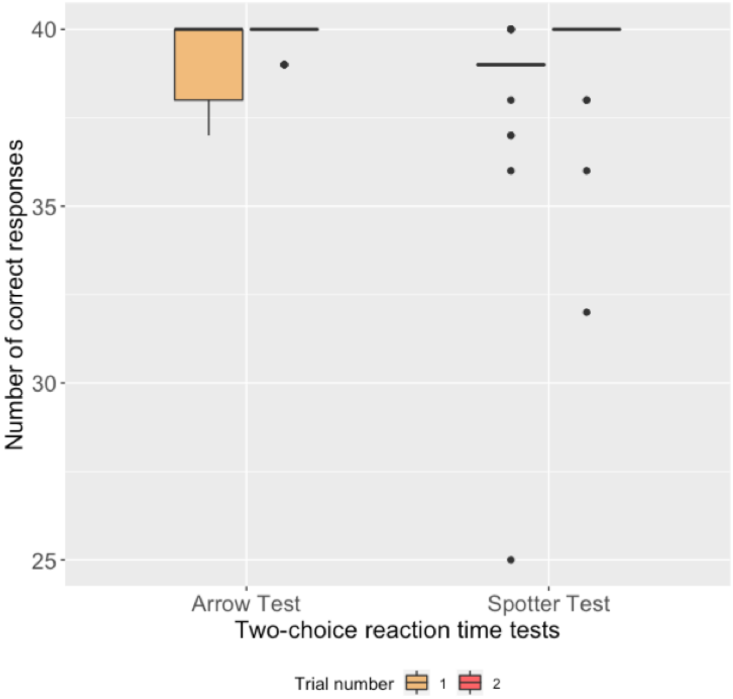


Figure 7. Box plots of participants' fastest response times during the two-choice reaction-time tests.

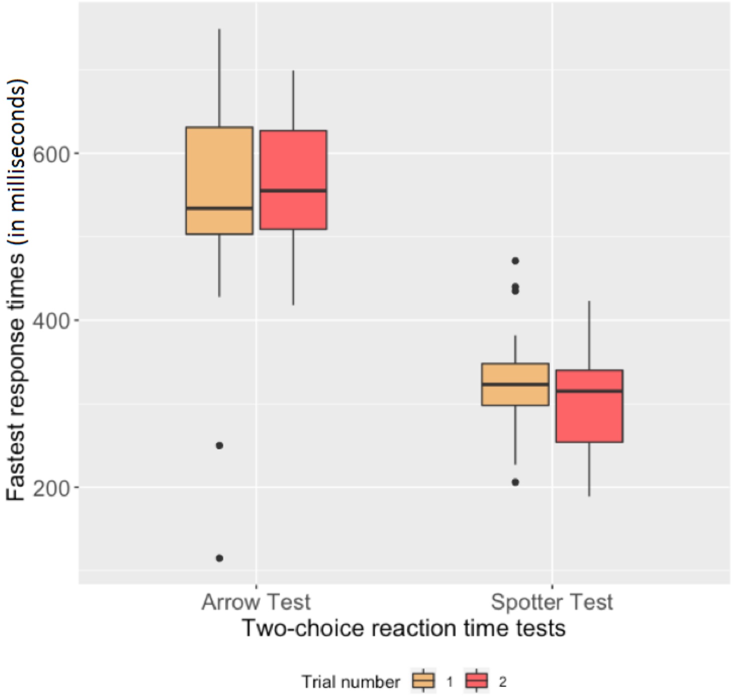


Figure 8. Box plots of participants' number of correct responses in the N-back tests.

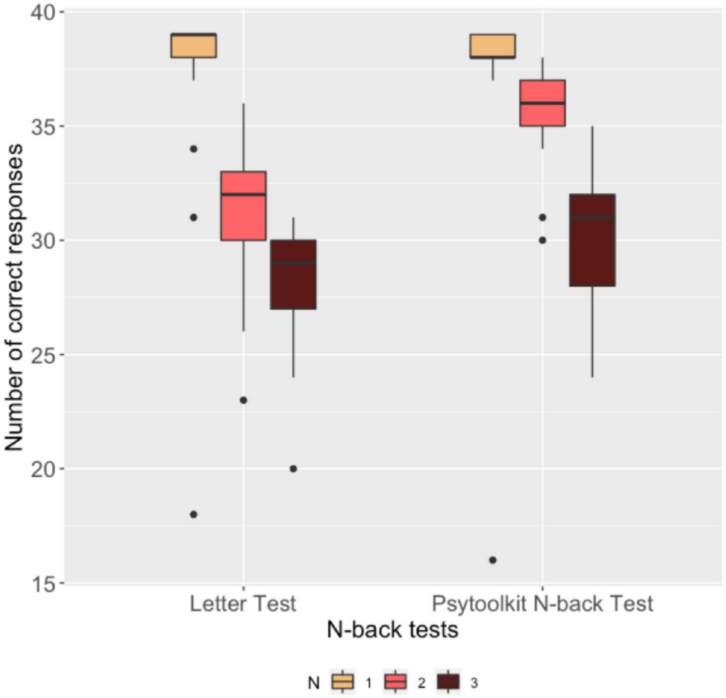


Figure 9. Box plots of participants' mean response times during the N-back tests.

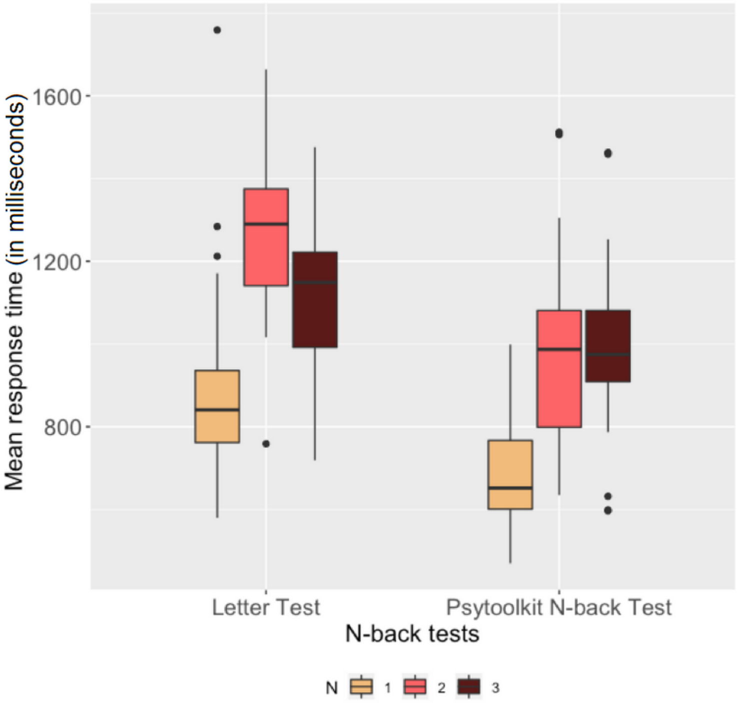


Table 2. Correlation analysis between performance measures of the N-back tasks in the Letter test and PsyToolkit N-back test.

Performance measure and tasks	Pearson *r*	*P* value
Mean response time		
1-back	.78	<.001
2-back	.71	<.001
3-back	.53	.01
Number of correct responses		
1-back	.90	<.001
2-back	.19	.40
3-back	.35	.13

One-way ANOVA was performed to analyze the effect of difficulty level on the participants' test performances (see Multimedia Appendix 2).

The results of the NASA-TLX questionnaire for 1-back, 2-back, and 3-back on both platforms are reported in Table 3. The numbers in this table show the means and SDs calculated based on the 7-point Likert scales for the metrics of the NASA-TLX.

Table 3. The N-back cognitive workload results using the NASA-TLX (Task Load Index) metrics.

Device and task	Score[a] for each metric, mean (SD)				
	Mental demand	Temporal demand	Overall performance	Effort	Frustration level
Smartwatch					
1-back	2.81 (1.50)	2.81 (1.29)	2.05 (1.32)	3.05 (1.20)	2.52 (1.60)
2-back	4.71 (1.27)	4.19 (1.50)	4.29 (1.49)	4.43 (1.17)	3.86 (1.80)
3-back	5.19 (1.33)	4.05 (1.75)	4.67 (1.62)	5.10 (1.10)	3.95 (1.80)
Computer					
1-back	2.76 (0.99)	2.86 (1.62)	2.91 (1.84)	2.67 (1.07)	2.48 (1.47)
2-back	4.50 (1.54)	3.50 (1.61)	3.10 (1.52)	4.35 (1.35)	2.95 (1.64)
3-back	5.52 (1.29)	4.24 (1.76)	4.76 (1.76)	5.00 (1.18)	4.00 (1.73)

[a] Scores were based on the 7-point Likert scales of the NASA-TLX metrics.

Stroop Color-Word Test

Figures 10 and 11 present the box plots of the mean RTs to the congruent and incongruent stimuli for each trial of the Color test and the PsyToolkit Stroop test, respectively. Table 4 reports the correlation analysis between the performance measures of the Stroop tests on both platforms. Box plots of the number of correct responses to both congruent and incongruent stimuli are shown in Figure 12.

Usability Ratings

The psychometric factors considered for the usability test were aesthetics, functionality, and information. Each of the UbiCAT apps were rated separately by the participants. Table 5 reports the means and SDs of the usability ratings, which are out of 5.

Figure 10. Box plots of participants' mean response times to congruent stimuli during Stroop tests.

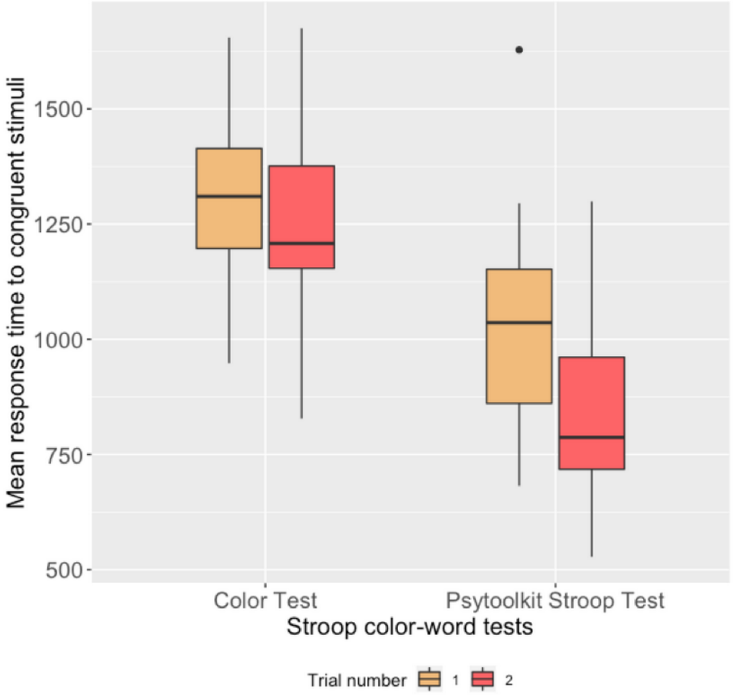


Figure 11. Box plots of participants’ mean response times to incongruent stimuli during the Stroop tests.

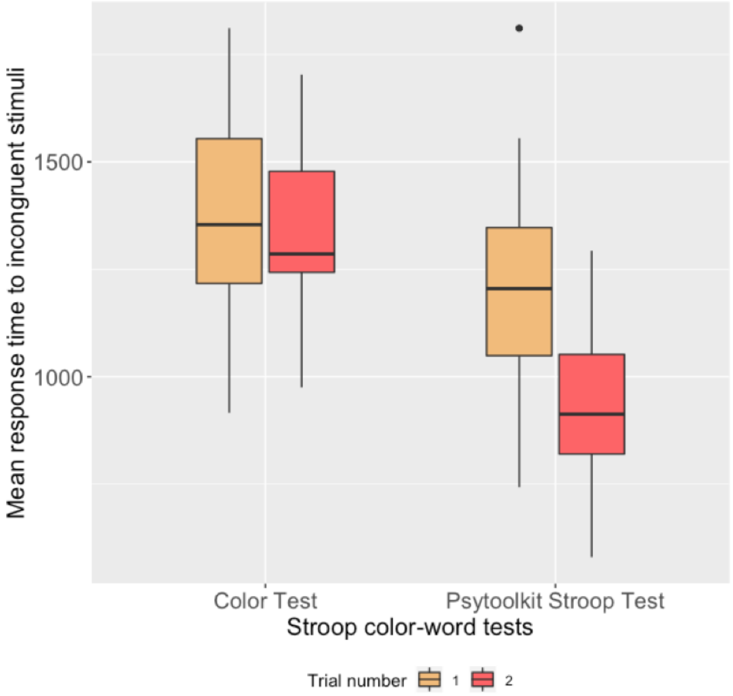


Table 4. Correlation analysis between performance measures in the Color test and the Stroop color-word test.

Performance measure	Pearson *r*	*P* value
Mean response times to congruent stimuli	.67	<.001
Mean response times to incongruent stimuli	.66	.001

Figure 12. Box plots of participants’ number of correct responses in the Stroop tests.

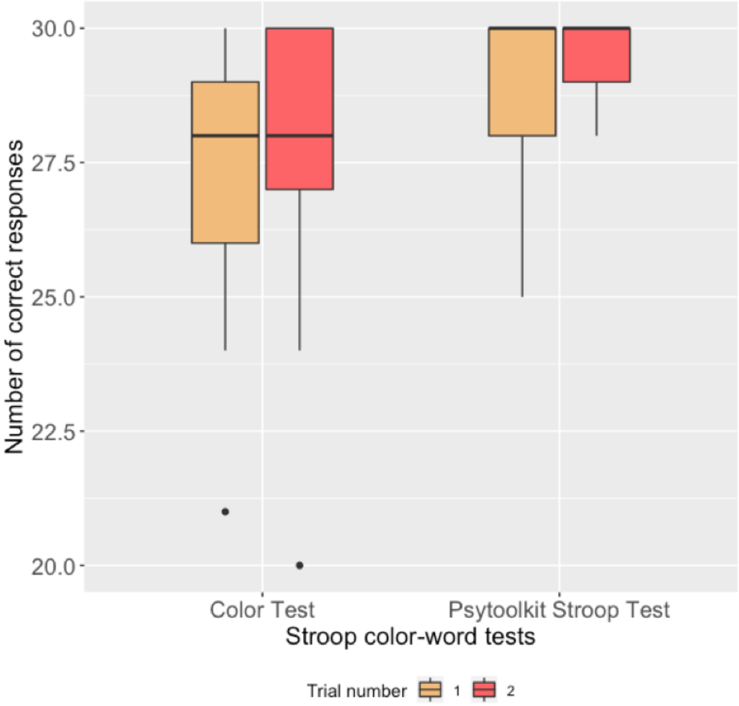


Table 5. Usability ratings of the UbiCAT (Ubiquitous Cognitive Assessment Tool) apps.

UbiCAT app	Score[a] for each factor, mean (SD)		
	Aesthetics	Functionality	Information
Arrow test	4.02 (0.76)	4.55 (0.52)	4.24 (0.86)
Letter test	4.19 (0.75)	4.36 (0.62)	4.33 (0.60)
Color test	4.14 (0.83)	4.64 (0.45)	4.31 (0.64)

[a]Scores ranged from 1 (the lowest score) to 5 (the highest score).

Perceived Discomfort

For each UbiCAT app, participants rated the discomfort level in their wrist on which they wore the smartwatch via a 7-point Likert scale from 1 (the least discomfort) to 7 (the most discomfort). The corresponding means and SDs of the discomfort levels, calculated separately for the Arrow test, Letter test, and Color test, are 2.71 (SD 1.79), 2.24 (SD 1.18), and 2.14 (SD 1.32), respectively.

Interviews

Overview

Seven themes were extracted from the participants' responses and a brief description of each theme is presented below. The participants’ quotes are presented in Multimedia Appendix 3.

Perceptions About the Experiment

Participants were asked to describe their feelings about the experiment. They were generally engaged in the experiment: 5 participants out of 21 (24%) mentioned that the experiment was “fun,” 3 (14%) said it was “good,” and 3 (14%) said it was “fine.” Only 1 participant out of 21 (5%) believed that the experiment was too long. The rest of the participants did not express their opinions or had to leave immediately after the experiment.

Input Modality

Participants compared the input modalities of the smartwatch and computer. Participants #1 and #3 (2/21, 10%) preferred the app buttons of the UbiCAT Color test to the keyboard in the Stroop test. Participants #9, #12, and #19 (3/21, 14%) felt more comfortable with the app buttons in general, and participant #14 (1/21, 5%) liked the tangibility of the keyboard.

Device Screen

Some participants compared the screen size of the smartwatch with the computer. Out of 21 participants, 1 (5%) argued that the bigger screen of the computer influenced his or her performance positively and 2 (10%) preferred the screen size of the computer to the smartwatch. We understood that computer screen size might be more acceptable for some people due to the longer adoption time of personal computers compared to smartwatches.

Visual Impact

Out of 21 participants, 3 (15%) implied that their better performance on the computer was due to the visualization of the elements. A participant (1/21, 5%) did not like the visual elements of the Fitbit, indicating that the overall device design and graphics affected participants' interaction quality apart from the specific user-interface design of the UbiCAT apps.

Psychological Factors

Apart from the physical characteristics of a smartwatch and a computer, psychological factors also influenced participants' performance. A participant (1/21, 5%) pointed to the gamified nature and playfulness of the UbiCAT tests.

Performance

Some of the participants related their lower performance in the UbiCAT tests to the apps. Out of 21 participants, 4 (19%) mentioned that the Color test sometimes did not capture their taps on the app buttons during the test. We noticed such an incident while reviewing the records of the experiments. The position of the app buttons in the Color test changed randomly to avoid practicing the positions of the buttons. It surprised some of the participants during the test. Besides, 1 of the participants (5%) thought that his or her performance might differ significantly between the first and second trials of the cognitive tests.

Suggestions

Of the 21 participants, 3 of them (14%) proposed suggestions regarding the font size used in the UbiCAT tests.

Discussion

Principal Findings

UbiCAT implements three smartwatch-based cognitive assessment tests for in-the-wild deployment. The findings of this study revealed comparable performance measures to computer-based tests. The strong correlation between the overall accuracy of the participants during the cognitive tests in the UbiCAT and computerized tools showed that UbiCAT can be utilized for assessing individuals' three key cognitive functions, namely attention, working memory, and executive function. The analysis between the following performance measures of the UbiCAT and computerized tests revealed significant correlation coefficients: the number of correct responses in the two-choice reaction-time test; mean RTs in the 1-back, 2-back, and 3-back tests; the number of correct responses in the 1-back test; and the mean RTs to the Stroop test’s congruent and incongruent stimuli.

The psychometric factors, including aesthetics, functionality, and information quality and quantity, of the UbiCAT apps had high average ratings by the participants (>4 out of 5). The subjective ratings of the participants' wrist discomfort levels were less than 3 out of 7, indicating that interaction with the UbiCAT apps via the smartwatch was comfortable, which is in line with our overall objective of making cognitive assessment as simple and convenient as possible.

Previous work reported mobile cognitive test results along with paper-based or computerized tests. Comparison between the correlation coefficients reported in previous studies and in our study is not possible due to different parameters, number of participants, and target population. Nevertheless, our test outcomes obtained from computer- and smartwatch-based apps were comparable to each other, unlike some of the previous studies (eg, Neurophone Stop Signal test) that could not compare their results with computerized or paper-based tests due to dissimilar parameters.

Two-Choice Reaction-Time Test Outcomes

The average number of correct responses obtained from the THINC-it and Arrow tests correlated significantly with each other. As it can be seen in Figure 6, the majority of the participants received the highest score on both platforms, which may indicate a ceiling effect. The participants' fastest RTs, however, did not correlate with each other, which might be due to the different interaction methods on both platforms. The app buttons in the Arrow test (see Figure 1) disappeared on receiving an input or time-out until the next stimulus appeared, since an accidental tap on the buttons could impede calculating the real performance of the participants. We observed that the participants moved their index fingers away after tapping on an app button in the Arrow test, while they kept their fingers on the arrow keys on the computer keyboard during the THINC-it Spotter test. Such a difference between the users' interactions may explain the longer RTs of the UbiCAT Arrow test as compared to the THINC-it Spotter test. Nevertheless, the difference between the fastest RTs of the participants helped us in understanding the impact of interaction methods.

The fastest RTs measured via both platforms may indicate that the thresholds of individuals' alertness vary on the computer and smartwatch platforms. In our study, the average fastest RTs of the participants in the Arrow test was 545 ms (SD 88), while the corresponding result for the THINC-it Spotter test was 315 ms (SD 59). A study on the development of a brief version of the PVT (PVT-B) showed that 500 ms might be the threshold for an impaired alertness [30], which is in line with the average fastest RTs obtained from the THINC-it computer-based test. However, the participants of the PVT-B study pressed a button to respond during the tests, which is similar to the interaction method of the THINC-it Spotter test. Therefore, this threshold may not be comparable to the fastest RTs obtained from the Arrow test on the smartwatch. To infer the level of impairment on the basis of the user's fastest RT delivered via a smartwatch, a larger study is required that would include both healthy controls and cognitively impaired patients. Nevertheless, the findings of our study revealed that the average fastest RT of the healthy subjects to a smartwatch-based test is above 500 ms.

According to Figure 7, participants' fastest RTs were almost the same during the first and second trials of the Spotter test (327 ms and 303 ms, respectively), while their responses were a bit slower in the second trial of the Arrow test (564 ms) compared to the first trial (526 ms). A paired-sample *t* test showed that the fastest RTs received from the Arrow test in both trials were not statistically different.

N-Back Test Outcomes

One-way ANOVA showed the effect of difficulty level on the number of correct responses and mean RTs in the UbiCAT Letter test and the PsyToolkit N-back test. The perceived cognitive workload in the N-back tests also revealed that as N-back tasks became more difficult, the participants' cognitive workload increased. The mean RTs obtained from each N-back task on both the smartwatch and the computer correlated significantly. Figure 9 shows that the mean RT during the UbiCAT 2-back test was higher than that of the 3-back test, while statistical analysis revealed no significant difference between the mean RTs of the UbiCAT 2-back and 3-back tests. The RTs of the PsyToolkit 2-back and 3-back tests were not statistically different either (*P*>.99). According to Table 3, higher temporal effort reported through NASA TLX questionnaires for the 2-back Letter test compared to the 3-back test may imply that participants were more rushed during the 2-back test. Moreover, participants might have spent more time on practicing the 2-back test right after taking the 1-back test to adapt their mental skills, since the reported mental effort for both 2-back and 3-back tests were higher than for the 1-back test on both the computer and the smartwatch.

According to Table 2, the correlation analysis between the number of correct responses of the N-back tests on the smartwatch and the computer was only significant for the 1-back test. The lack of a significant correlation between the 2-back and 3-back tasks might be due to the N-back test itself, since the letter sequences of the N-back test were generated randomly and the maximum number of matches (ie, hits) during the N-back tests was not controlled to be the same between the computer and the smartwatch.

Stroop Test Outcomes

The RTs to the congruent and incongruent stimuli on the PsyToolkit and Color tests significantly correlated with each other. According to Figures 10 and 11, the RTs obtained from the second trials were lower than the first trials for both the PsyToolkit Stroop test and Color test. However, the magnitude of difference between the RTs in both trials of the Color test was lower than that in the PsyToolkit Stroop test. It might be due to the difference between the interaction methods of the tests. In the Color test, the order of app buttons was shuffled after a test run to avoid practicing the positions and increasing engagement with the apps. On the other hand, the position of the keys was obviously stable during the PsyToolkit tests. Hence, participants might get used to the position of the keys and respond faster in the second trial of the PsyToolkit Stroop test, while the changing position of the app buttons in the Color test took some time for them to practice with the new positions. The change in the position of the app buttons was intended to obtain reliable outcomes during future studies for longitudinal frequent administration.

Figure 12 shows that several participants received the highest score in the PsyToolkit Stroop tests (ie, 10 participants in trial 1 and 11 participants in trial 2), while the scores are more distributed in the Color tests (ie, 2 participants received the highest score in trial 1 and 6 participants in trial 2). In addition, we observed that sometimes the app buttons in the Color test did not capture touch inputs by the participants and some participants reported this issue during the interviews. Therefore, lower scores in the Color test might be due to the Fitbit's touch sensitivity.

Perceptions From the Interviews

Seven themes were identified from the follow-up interviews with the participants. Some of the participants generally felt more comfortable when taking a cognitive test on the smartwatch compared to the computer, while some did not. Factors related to the physical aspects of the device, including the screen size and distance and the input modalities, affected their interactions. We understood that longer adoption times of computers compared to smartwatches may explain why some participants preferred computer tests to the UbiCAT apps. Therefore, deploying smartwatches into individuals' daily lives may take some time and may not be useful for all. Psychological factors were also involved in determining participants' engagement with UbiCAT, such as the gamified features of the tests.

Implications for Future Work

On the basis of our interviews, we decided to (1) add customized badges to the UbiCAT apps depending on participants' test performances to motivate them toward continuous usage of the UbiCAT, (2) increase the font size of the stimulus in the Letter test since it was not easy for some of the participants to read, and (3) keep the right and left app buttons of the Arrow test on the screen after they tap on a button.

This study was conducted to evaluate our novel smartwatch apps against their corresponding computer tests, as well as to investigate participants' perceptions about the study and usability of the UbiCAT apps. One of the future directions of the UbiCAT project is to identify digital biomarkers of human cognition. In an upcoming study, we will collect participants' mobile data, including physiological and behavioral data, along with assessing their daily cognitive functioning through the UbiCAT apps to determine digital biomarkers of human cognition. The digital biomarkers would help researchers in building predictive models of individuals' cognitive impairment using mobile data.

Sleep-stage data and heart rate variability (HRV) are the physiological data that can be collected through the Fitbit API. Sleep disturbance is, for instance, prevalent in bipolar patients [42]. The negative impact of poor sleep on mood and cognitive functioning is particularly noticeable in bipolar patients [43]. Fitbit smartwatches collect sleep duration and stages, which can help us in measuring the impact of sleep quality on next-day cognitive performance. Literature suggests a relationship between reduced HRV and impairment in inhibition control [44]. Moreover, reduced HRV was observed in bipolar and schizophrenic patients compared with healthy controls [45]. Our outlook is to create a *Cognitive Watch* to extend human knowledge regarding cognition.

Limitations

The limitations of this study are threefold. First, this study was conducted with 21 healthy adults who were recruited mostly from the campus of the Technical University of Denmark. This was deemed appropriate for evaluating the UbiCAT as compared to existing tools. However, studies involving patients and people with cognitive impairment are needed and are the focus of our upcoming studies. Second, the UbiCAT is designed for in-the-wild administration and, yet, this study was conducted in an indoor environment. This was done because cognitive performance fluctuates and in order to be able to assess cognition using both UbiCAT and the computer-based tests, the tests had to be administered right after each other on both platforms in order to achieve comparable measures. Therefore, moving the participants inside and outside between computer-based and smartwatch-based test sessions could yield unreliable cognitive measures for our study. In our upcoming studies, however, the UbiCAT will be used outside the clinic in order to collect real-world cognitive performance, which will be compared with cognitive assessments performed in a clinic. Third, the results indicated that the UbiCAT tests may not reflect the optimal performance of the participants compared to the computer-based tests. Nevertheless, frequent tests with the UbiCAT in upcoming studies and with various patient groups may better verify the optimal performance of the UbiCAT users.

Conclusions

In this study, the UbiCAT as a smartwatch-based tool for cognitive assessment was evaluated against computer-based cognitive assessment tools. The results revealed significant correlations between the total scores of the UbiCAT tests and standard computer-based tests. The psychometric factors regarding the aesthetics, functionality, and information quality and quantity of the apps yielded high usability ratings from the study participants. The majority of our study participants felt comfortable when using the UbiCAT. The findings of this study showed that the UbiCAT can be used for assessing attention, working memory, and executive function across participants' everyday lives, along with mobile data collection. Future studies can administer the UbiCAT to mentally ill patients to collect their daily cognitive functioning data and to compare their results with lab-based studies.

Acknowledgments

This research is funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie grant agreement (No. 722561). We wish to thank our study participants.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Usability evaluation questionnaire.
[PDF File (Adobe PDF File), 68 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Analysis of variance on the N-back test results.
[PDF File (Adobe PDF File), 30 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Themes and their corresponding quotes extracted from the interviews.
[PDF File (Adobe PDF File), 47 KB-Multimedia Appendix 3]

References

1. Khakurel J, Immonen M, Porras J, Knutas A. Understanding the adoption of quantified self-tracking wearable devices in the organization environment. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19). 2019 Presented at: 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '19); June 5-7, 2019; Rhodes, Greece. [doi: 10.1145/3316782.3321527]

2. Chuah SH, Rauschnabel PA, Krey N, Nguyen B, Ramayah T, Lade S. Wearable technologies: The role of usefulness and visibility in smartwatch adoption. Comput Human Behav 2016 Dec;65:276-284. [doi: 10.1016/j.chb.2016.07.047]
3. King CE, Sarrafzadeh M. A survey of smartwatches in remote health monitoring. J Healthc Inform Res 2018 Jun;2(1-2):1-24 [FREE Full text] [doi: 10.1007/s41666-017-0012-7] [Medline: 30035250]
4. Lyon GR, Krasnegor NA. Attention, Memory, and Executive Function. Baltimore, MD: Paul Brookes Publishing Co; 1995.
5. Schmidt C, Collette F, Cajochen C, Peigneux P. A time to think: Circadian rhythms in human cognition. Cogn Neuropsychol 2007 Oct;24(7):755-789. [doi: 10.1080/02643290701754158] [Medline: 18066734]
6. Folkard S. Time of day and level of processing. Mem Cognit 1979 Jul;7(4):247-252 [FREE Full text] [doi: 10.3758/bf03197596]
7. Lara T, Madrid JA, Correa Á. The vigilance decrement in executive function is attenuated when individual chronotypes perform at their optimal time of day. PLoS One 2014;9(2):e88820 [FREE Full text] [doi: 10.1371/journal.pone.0088820] [Medline: 24586404]
8. Lee D, Taylor J, Thomas A. Assessment of cognitive fluctuation in dementia: A systematic review of the literature. Int J Geriatr Psychiatry 2012 Oct;27(10):989-998. [doi: 10.1002/gps.2823] [Medline: 22278997]
9. Bora E, Özerdem A. Meta-analysis of longitudinal studies of cognition in bipolar disorder: Comparison with healthy controls and schizophrenia. Psychol Med 2017 Dec;47(16):2753-2766. [doi: 10.1017/S0033291717001490] [Medline: 28585513]
10. Szmulewicz A, Valerio MP, Martino DJ. Longitudinal analysis of cognitive performances in recent-onset and late-life bipolar disorder: A systematic review and meta-analysis. Bipolar Disord 2020 Feb;22(1):28-37. [doi: 10.1111/bdi.12841] [Medline: 31541587]
11. Fuermaier AB, Tucha L, Koerts J, Aschenbrenner S, Kaunzinger I, Hauser J, et al. Cognitive impairment in adult ADHD: Perspective matters!. Neuropsychology 2015 Jan;29(1):45-58. [doi: 10.1037/neu0000108] [Medline: 24933488]
12. Elvevåg B, Goldberg TE. Cognitive impairment in schizophrenia is the core of the disorder. Crit Rev Neurobiol 2000;14(1):1-21. [Medline: 11253953]
13. Cambridge Cognition. CANTAB Mobile URL: http://www.cambridgecognition.com/products/digital-healthcare-technology/cantab-mobile/ [accessed 2019-01-08] [WebCite Cache ID 75GimIcUD]
14. Hafiz P, Miskowiak KW, Kessing LV, Elleby Jespersen A, Obenhausen K, Gulyas L, et al. The internet-based cognitive assessment tool: System design and feasibility study. JMIR Form Res 2019 Jul 26;3(3):e13898 [FREE Full text] [doi: 10.2196/13898] [Medline: 31350840]
15. Harrison JE, Barry H, Baune BT, Best MW, Bowie CR, Cha DS, et al. Stability, reliability, and validity of the THINC-it screening tool for cognitive impairment in depression: A psychometric exploration in healthy volunteers. Int J Methods Psychiatr Res 2018 Sep;27(3):e1736 [FREE Full text] [doi: 10.1002/mpr.1736] [Medline: 30088298]
16. Domen AC, van de Weijer SC, Jaspers MW, Denys D, Nieman DH. The validation of a new online cognitive assessment tool: The MyCognition Quotient. Int J Methods Psychiatr Res 2019 Sep;28(3):e1775 [FREE Full text] [doi: 10.1002/mpr.1775] [Medline: 30761648]
17. Maruff P, Thomas E, Cysique L, Brew B, Collie A, Snyder P, et al. Validity of the CogState brief battery: Relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. Arch Clin Neuropsychol 2009 Mar;24(2):165-178. [doi: 10.1093/arclin/acp010] [Medline: 19395350]
18. Keefe RS, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L. The Brief Assessment of Cognition in Schizophrenia: Reliability, sensitivity, and comparison with a standard neurocognitive battery. Schizophr Res 2004 Jun 01;68(2-3):283-297. [doi: 10.1016/j.schres.2003.09.011] [Medline: 15099610]
19. Timmers C, Maeghs A, Vestjens M, Bonnemayer C, Hamers H, Blokland A. Ambulant cognitive assessment using a smartphone. Appl Neuropsychol Adult 2014;21(2):136-142. [doi: 10.1080/09084282.2013.778261] [Medline: 24826507]
20. Tiplady B, Oshinowo B, Thomson J, Drummond GB. Alcohol and cognitive function: Assessment in everyday life and laboratory settings using mobile phones. Alcohol Clin Exp Res 2009 Dec;33(12):2094-2102. [doi: 10.1111/j.1530-0277.2009.01049.x] [Medline: 19740132]
21. Stone AA, Shiffman S. Ecological Momentary Assessment (EMA) in behavioral medicine. Ann Behav Med 1994 Sep 01;16(3):199-202. [doi: 10.1093/abm/16.3.199]
22. Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. In: Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi. Dordrecht, the Netherlands: Springer; 2014:35-54.
23. Trull TJ, Ebner-Priemer UW. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. Psychol Assess 2009 Dec;21(4):457-462 [FREE Full text] [doi: 10.1037/a0017653] [Medline: 19947780]
24. Moore RC, Swendsen J, Depp CA. Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. Int J Methods Psychiatr Res 2017 Dec;26(4):1-12 [FREE Full text] [doi: 10.1002/mpr.1562] [Medline: 28370881]
25. Jongstra S, Wijsman LW, Cachucho R, Hoevenaar-Blom MP, Mooijaart SP, Richard E. Cognitive testing in people at increased risk of dementia using a smartphone app: The iVitality proof-of-principle study. JMIR Mhealth Uhealth 2017 May 25;5(5):e68 [FREE Full text] [doi: 10.2196/mhealth.6939] [Medline: 28546139]

26. Brouillette RM, Foil H, Fontenot S, Correro A, Allen R, Martin CK, et al. Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. PLoS One 2013;8(6):e65925 [FREE Full text] [doi: 10.1371/journal.pone.0065925] [Medline: 23776570]
27. Tieges Z, Stíobhairt A, Scott K, Suchorab K, Weir A, Parks S, et al. Development of a smartphone application for the objective detection of attentional deficits in delirium. Int Psychogeriatr 2015 Aug;27(8):1251-1262. [doi: 10.1017/S1041610215000186] [Medline: 25742756]
28. Pal R, Mendelson J, Clavier O, Baggott MJ, Coyle J, Galloway GP. Development and testing of a smartphone-based cognitive/neuropsychological evaluation system for substance abusers. J Psychoactive Drugs 2016;48(4):288-294. [doi: 10.1080/02791072.2016.1191093] [Medline: 27260123]
29. Dingler T, Schmidt A, Machulla T. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Sep 11;1(3):1-15 [FREE Full text] [doi: 10.1145/3132025]
30. Basner M, Mollicone D, Dinges DF. Validity and sensitivity of a Brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation. Acta Astronaut 2011 Dec 01;69(11-12):949-959 [FREE Full text] [doi: 10.1016/j.actaastro.2011.07.015] [Medline: 22025811]
31. Cormack F, McCue M, Taptiklis N, Skirrow C, Glazer E, Panagopoulos E, et al. Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: Longitudinal observational study. JMIR Ment Health 2019 Nov 18;6(11):e12814 [FREE Full text] [doi: 10.2196/12814] [Medline: 31738172]
32. Donders F. On the speed of mental processes. Acta Psychol 1969;30:412-431 [FREE Full text] [doi: 10.1016/0001-6918(69)90065-1]
33. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol 1935;18(6):643-662. [doi: 10.1037/h0054651]
34. Kirchner WK. Age differences in short-term retention of rapidly changing information. J Exp Psychol 1958 Apr;55(4):352-358. [doi: 10.1037/h0043688] [Medline: 13539317]
35. Hafiz P, Bardram JE. Design and formative evaluation of cognitive assessment apps for wearable technologies. In: Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers (UbiComp/ISWC'19 Adjunct). 2019 Presented at: ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers (UbiComp/ISWC'19 Adjunct); September 9-13, 2019; London, UK p. 1162-1165 URL: https://dl.acm.org/doi/pdf/10.1145/3341162.3347077 [doi: 10.1145/3341162.3347077]
36. Stoet G. PsyToolkit: A software package for programming psychological experiments using Linux. Behav Res Methods 2010 Nov;42(4):1096-1104. [doi: 10.3758/BRM.42.4.1096] [Medline: 21139177]
37. Stoet G. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. Teach Psychol 2017;44(1):24-31. [doi: 10.1177/0098628316677643]
38. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd edition. Berlin, Germany: Springer Nature; 2016.
39. Biernacki P, Waldorf D. Snowball sampling: Problems and techniques of chain referral sampling. Sociol Methods Res 2016 Jun 29;10(2):141-163. [doi: 10.1177/004912418101000205]
40. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: A new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
41. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Adv Psychol 1988;52:139-183. [doi: 10.1016/s0166-4115(08)62386-9]
42. Harvey AG, Talbot LS, Gershon A. Sleep disturbance in bipolar disorder across the lifespan. Clin Psychol (New York) 2009 Jun;16(2):256-277 [FREE Full text] [doi: 10.1111/j.1468-2850.2009.01164.x] [Medline: 22493520]
43. Ancoli-Israel S, Roth T. Characteristics of insomnia in the United States: Results of the 1991 National Sleep Foundation Survey. I. Sleep 1999 May 01;22 Suppl 2:S347-S353. [Medline: 10394606]
44. Thayer JF, Lane RD. Claude Bernard and the heart-brain connection: Further elaboration of a model of neurovisceral integration. Neurosci Biobehav Rev 2009 Feb;33(2):81-88. [doi: 10.1016/j.neubiorev.2008.08.004] [Medline: 18771686]
45. Henry BL, Minassian A, Paulus MP, Geyer MA, Perry W. Heart rate variability in bipolar mania and schizophrenia. J Psychiatr Res 2010 Feb;44(3):168-176 [FREE Full text] [doi: 10.1016/j.jpsychires.2009.07.011] [Medline: 19700172]

Abbreviations

ADHD: attention deficit hyperactivity disorder
ANOVA: analysis of variance
API: application programming interface
BACS: Brief Assessment of Cognition in Schizophrenia
CANTAB: Cambridge Neuropsychological Test Automated Battery
CST: Color-Shape Test
EDTB: Edinburgh Delirium Test Box

EMA: ecological momentary assessment
ESM: experience-sampling method
HRV: heart rate variability
ICAT: Internet-based Cognitive Assessment Tool
MARS: Mobile App Rating Scale
MyCQ: MyCognition Quotient
NASA TLX: NASA Task Load Index
OS: operating system
PVT: psychomotor vigilance task
PVT-B: brief version of the psychomotor vigilance task
RT: response time
THINC-it: THINC-integrated tool
UbiCAT: Ubiquitous Cognitive Assessment Tool
UDS: Uniform Data Set

Edited by E Borycki; submitted 17.12.19; peer-reviewed by A Blokland, J Drijver; comments to author 02.02.20; revised version received 10.03.20; accepted 31.03.20; published 01.06.20

Please cite as:
Hafiz P, Bardram JE
The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study
JMIR Mhealth Uhealth 2020;8(6):e17506
URL: https://mhealth.jmir.org/2020/6/e17506
doi: 10.2196/17506
PMID:

10.3 Analysis of Perceived Human Factors and Participants’ Demographics during a Cognitive Assessment Study with a Smartwatch

Authors: **Pegah Hafiz**, Alban Maxhuni, Jakob E Bardram

Accepted in the 8th IEEE International Conference on Healthcare Informatics

Analysis of Perceived Human Factors and Participants' Demographics during a Cognitive Assessment Study with a Smartwatch

Pegah Hafiz
Department of Health Technology
Technical University of Denmark
Kongens Lyngby, Denmark
pegh@dtu.dk

Alban Maxhuni
Department of Health Technology
Technical University of Denmark
Kongens Lyngby, Denmark
almax@dtu.dk

Jakob E. Bardram
Department of Health Technology
Technical University of Denmark
Kongens Lyngby, Denmark
jakba@dtu.dk

***Abstract*—Digital tools have been developed to assess human cognitive functioning. It is unknown to what degree users' cognitive test performance is correlated with their perceived usability and cognitive load induced by interaction with a tool. Moreover, the similarity between user groups in terms of their subjective usability and cognitive load has not been explored adequately despite its potential importance in designing digital cognitive assessment tools for people from diverse background. This paper presents a study of two smartwatch-based cognitive tests to assess participants' attention and working memory. NASA Task Load Index (NASA-TLX) and Mobile App Rating Scale (MARS) questionnaires were used for cognitive load and usability evaluations, respectively. Aesthetics, functionality, and information quality and quantity were the metrics we selected for usability evaluations. Pearson's correlation analysis was performed to investigate the associations and Ward's clustering method was applied for data visualization. Our results showed that participants who received higher scores and longer scoring streak rated functionality of the cognitive tests better. Moreover, information quality and quantity of the tests were rated better by the participants who received longer scoring streak indicating the significant role of test instructions in gaining higher scores. In addition, participants with lower temporal demand received higher scores and faster mean response times. The key findings from the clusters visualized in this paper are: (i) Female and male participants rated their perceived usability and cognitive load completely differently; (ii) A discrepancy was found between participants' perceived performance and their actual scores; (iii) Participants from diverse background rated their perceived usability and cognitive load different from each other.**

***Index Terms*—cognition, cognitive load, usability, human factor, correlation, working memory, attention, clustering**

I. Background

USABILITY and cognitive load metrics are two major constructs of human factors that have been studied in many domains. The International Organisation for Standardisation (ISO 9241) [1] defines usability metrics as effectiveness, efficiency, and satisfaction. A recent review investigated the usability methods of mHealth applications and found that approximately 50% of the studies (13 out of 27) used questionnaires and evaluated psychometric factors including attractiveness, learnability, operability, and understandability of the applications [2]. Cognitive load is another crucial aspect of human factors since excessive mental workload induced by an application can lead to a negative impact on users' learnability [3]–[5].

Cognitive functioning is a key aspect of human mental health. An impairment in attention, memory, and executive function can cause problems for individuals at their work or school [6]. The tests for assessing cognitive functioning often put mental pressure on the users' brain. Digital tools for cognitive assessment have been designed for personal computers, tablets, and mobile devices. Examples include Cambridge Neuropsychological Test Automated Battery (CANTAB) Mobile [7], the THINC- Integrated Tool (THINC-it) [8], CogState [9], and the Internet-based Cognitive Assessment Tool (ICAT) [10]. These tools provide remote assessment of both healthy individuals and patients. Recently, smartwatch-based tools have emerged that allow for *'in-the-wild'* assessments. Examples include the Cognition Kit [11] for assessment of working memory and the Ubiquitous Cognitive Assessment Tool (UbiCAT) [12] for assessment of alertness, working memory, and executive function.

Overall, digital cognitive assessment tools have gained momentum and are administered in a wide range of user groups from diverse backgrounds. Although these tools have shown promising feasibility, some issues are introduced by them. First, usability may play a significant role in such tests and it is essential to determine whether users are able to take the tests properly via the user interfaces. Inability to interact with a tool can potentially impact the assessment of the users' real cognitive functioning. Second, cognitive load induced by digital cognitive assessment tools may similarly impact the users taking the tests, which again may negatively affect their test results. As such, usability and cognitive load are the factors that might influence assessment of the users' real cognitive functioning. Third, participants' demographics are often collected in surveys in which human factors such as usability and mental workload are also assessed. Descriptive statistics of participants' demographics are often reported while little attention is paid to the similarity or association between various participants in terms of their perceived usability and cognitive load after taking cognitive tests. Investigating the latter would tell about the design of a digital cognitive assessment tool for users with a focus on their demographics. In order to address these issues, this paper presents a study of a digital cognitive assessment tool, and seeks to investigate the following questions:

- What is the relationship between users' subjective usability and cognitive load metrics on the one hand, and their objective cognitive test performance on the other?
- How does users' demographic background (e.g., gender and education) relate to how they rate usability and cognitive load after taking a cognitive test with inherent mental pressure?

Answering these questions will help design more reliable as well as more usable digital cognitive assessment tools.

II. Related Work

Table I gives an overview of related work. A study identified the relationship between the usability of a website and personal factors, Intelligence Quotient (IQ) and cognitive abilities of students [13]. According to their finding, participants with higher IQ and Grade Point Average (GPA) rated the learnability of a software higher. Another experiment with students assessed the effect of system and user features on perceived usability and ease of use of a Web-based learning system [14]. The user features included subjective norm, self-efficacy, and innovativeness in information technology and system features involved computer playfulness, interface style, and interactivity. Their findings showed that the effect of user features was higher than system features on perceived usability while the impact of system features was higher than user features on the ease of use.

Van et al. conducted a study to find a relationship between usability of an internet-based cognitive behavioural therapy program for chronic pain and participants' sociodemographics [15]. Their findings revealed that usability negatively correlated with age and positively correlated with digital health knowledge while no correlation was found between usability and educational level. Some previous studies conducted with System Usability Scale (SUS) questionnaire [16] showed no impact of gender on their overall usability ratings [17]–[20]. Kortum and Oswald [21] evaluated usability of 14 frequently-used products using the SUS questionnaire. Their findings showed higher overall usability ratings in female participants regarding Word and Amazon products while the rest of the applications were rated higher by male participants. The authors in [22] evaluated usability of mobile banking apps and performed statistical analysis between users' satisfaction and their demographics. Their results showed that male participants were more satisfied with the mobile apps. Furthermore, participants at a Ph.D. level felt more content with the apps compared to the individuals in Master's and first-degree levels.

To our knowledge, none of the existing studies have explored the association between individuals' cognitive test performance delivered via a digital tool and their perceived usability and cognitive load metrics. In addition, the role of users' characteristics in usability and cognitive load assessment studies have not been adequately explored during cognitive assessment tests. The aforementioned gaps in the literature motivated us in setting the following objectives for the present work:

Table (I) Main related works in perceived human factor analysis showing the features used and details about the studies.

Study	Items measured	Evaluated system	Method
Karahoca et al. [13]	IQ; GPA	Web portal	Software usability measurement inventory
Ke et al. [14]	User features: subjective norm, self-efficacy, and innovativeness; System features: computer playfulness, interface style, and interactivity	Web-based learning system	Questionnaire (5-point Likert scale)
Van et al. [15]	Demographics: age, digital health knowledge, education level	Internet-based cognitive behavioural therapy	Num. of completed performance tasks; Num. of encountered problems

Kortum et al. [21]	Usability;
	Personality;
	Demographics: gender
		Frequently-used softwares/products	SUS

Mkpojiogu et al. [22]	Usability; Demographics: age, gender, education, experience	Mobile banking apps	Questionnaire (9-point Likert scale)

- To identify the correlation between individuals' cognitive test performance delivered via a digital tool and their perceived usability
- To identify the correlation between individuals' working memory performance and their perceived cognitive load
- To investigate similarities between the perceived usability and cognitive load measures of our study participants on the basis of their demographics

III. Methodology

The study protocol was sent for approval at the Danish Ethical Committee and was exempted from ethical approval as it was not a clinical survey (Journal-nr.: H-19086232). Participants were recruited on voluntary basis and an informed consent form was signed by them prior to the study. The participants' age, education level, and industry were collected as well as their cognitive test performance and subjective usability and cognitive load. We used two smartwatch-based apps of UbiCAT [12] and collected associated cognitive performance data. The apps in UbiCAT are short, engaging, and run on Fitbit Ionic smartwatches. This tool includes digital versions of Two-Choice Reaction Time (2-CRT) [23] and

N-back [24] tests which we used in our study. Three test performance measures including mean Response Time (RT), number of correct responses, and longest scoring streak were calculated by UbiCAT cognitive tests. Longest scoring streak is the maximum number of stimuli to which participants responded correctly without leaving any incorrect or missed response in between. The tests were timed which means users had limited time to respond to each test stimuli. It took approximately two min per participant to take each of the 2-CRT and 1-back tests. A snapshot of a study participant who took a 2-CRT test is presented in Figure 1.

Figure (1) A snapshot of the 2-CRT test in UbiCAT

The experiments were performed in a silent room. Participants wore a Fitbit smartwatch on their non-dominant hand. Each participant took the 2-CRT test for two consecutive trials to achieve a reliable measure of alertness and 1-back for one trial. Participants could check their scores at the end of a tests session.

Each test was followed by a usability questionnaire (see Appendix A). We selected seven questions from a validated usability tool called Mobile App Rating Scale (MARS) questionnaire [25]. The factors considered for usability are aesthetics, functionality, and information quantity and quality of the two apps presenting standard 2-CRT and 1-back tests. Furthermore, selected factors from the MARS questionnaire are inline with the frequent measures evaluated for mHealth apps [2].

Participants additionally rated their perceived cognitive load upon finishing the 1-back test in UbiCAT. It should be noted that N-back is a valid cognitive test that not only measures working memory but also have been utilized in several studies in which cognitive load of the individuals were measured (for example, [26]–[29]). NASA Task Load Index (NASA-TLX) questionnaire [30] was used to measure perceived cognitive load of the participants. We excluded a sub-scale of NASA-TLX regarding physical effort as it was not relevant to our study instrument. Hence, the sub-scales considered for the present study are mental demand, temporal demand, overall performance, effort, and frustration level.

Figure 2 illustrates the procedure of the first part of this paper where correlation analysis significant at 95% level was performed. In the second part, we applied Ward's method [31] as a hierarchical clustering technique to visualize participants' perceived usability and cognitive load metrics based on their demographics. The Ward's method uses half-square euclidean distance[1] between participants as presented in Equation (1). Finally, we grouped similarities of our study participants by gender, education level, and work/study industry using Equation (2). The Ward's method provides several advantages over other clustering algorithms: (i) There is no need to define the number of clusters for the algorithm; (ii) It is easy to implement; (iii) Dendrograms are useful in understanding the similarities.

$$\textbf{distance}(a_i, b_i) = \frac{1}{2}\sum_{i=1}^{k}(a_i - b_i)^2 \tag{1}$$

$$\textbf{similarity}_{(C1,C4)} = a\cdot sim_{(C1,C2)} + b\cdot sim_{(C1,C3)} - c\cdot sim_{(C2,C3)} \tag{2}$$

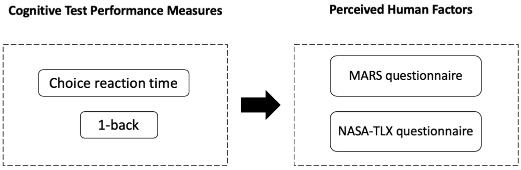


Figure (2) Schematic overview of the correlation analysis performed in this paper

IV. Results

In total, N=21 participants in Copenhagen, Denmark were selected for this study. Table II provides a summary of participant's demographic characteristics and it can be noted that there is a fairly balanced mix of gender (9-female, 12-male), education, age (M = 26.9, SD = 5.98), industry, and jobs among participants. Cognitive tests performance of the study participants are reported in Table III, where 2-CRT performance measures are calculated by averaging the values obtained from two consecutive trials. Usability ratings of the apps presenting 2-CRT and 1-back tests in UbiCAT are shown in Figure 3. Mean and standard deviations of the participants' perceived cognitive load for each sub-scale are depicted in Figure 4.

A. Correlation between Usability Metrics and Cognitive Test Measures

Table IV and Table V show the correlation coefficients between the usability metrics and the test performance measures for 2-CRT and 1-back tests, respectively. Strong correlation coefficients were revealed between participants' perceived functionality and achieved scores and longest scoring streaks in both 2-CRT and 1-back tests. It can be inferred that participants who received higher scores and achieved longer scoring streak rated functionality of the tests higher. In the 1-back test, the participants who were faster in responding rated

[1] Euclidean distance is always greater than or equal to zero. Measurements would be ≈ 0 for identical subjects and ≈ 1 for subjects that show less similarity.

Table (II) Study demographics of our participants.

Variable	Characteristics	Nr. (%)
Gender	Male	**12** (57.14%)
	Female	**9** (42.86%)
Education	Bachelor degree	**6** (28.57%)
	Master degree	**8** (38.10%)
	Ph.D.	**7** (33.33%)
Age	19-30	**17** (80.95%)
	31-40	**3** (14.29%)
	> 40	**1** (4.76%)
	Mean ± SD	**26.90 ± 5.98**
Industry	Design	**4** (19.05%)
	Research	**4** (19.05%)
	Computer Engineer	**4** (19.05%)
	Construction	**1** (4.76%)
	Education	**1** (4.76%)
	Energy Engineer	**1** (4.76%)
	Food Engineer	**1** (4.76%)
	Healthcare	**3** (14.29%)
	Research	**4** (19.05%)
	Water Engineer	**2** (9.52%)
Job	Student Assistant	**3** (14.29%)
	Bachelor Student	**3** (14.29%)
	Master Student	**5** (23.80%)
	Ph.D Student	**4** (19.05%)
	Postdoctoral Researchers	**3** (14.29%)
	Data Analyst	**1** (4.76%)
	Nurse	**1** (4.76%)
	Project Manager	**1** (4.76%)

Table (III) Mean and standard deviations of the participants' cognitive test performance during the choice reaction time and 1-back tests

Test	Response time	Correct responses	Longest streak
2-CRT	773±107	39.57±0.60	36.5±5.34
1-Back	903±266	37.09±4.82	34±9.86

the functionality better. Participants' longest scoring streak also correlated significantly with their perceived information quantity and quality.

B. Correlation between 1-Back Test Measures and Cognitive Load Sub-scales

Correlation analysis was applied between the sub-scales of NASA-TLX questionnaire, which was rated by the participants and the performance measures of 1-back tests. Significant correlation coefficients are reported as follows:

Mean RTs of the participants correlated moderately with their temporal demand (r= 0.54, p= 0.011) and effort (r= 0.50, p= 0.02). Number of correct responses correlated moderately with temporal demand (r= -0.45, p= 0.04) and frustration level (r= -0.47, p= 0.03). Similarly, the longest scoring streak of the participants correlated with temporal demand (r= -0.55, p= 0.009) and frustration level (r= -0.44, p= 0.04).

C. Correlation between Usability and Cognitive Load metrics

An analysis was performed between cognitive load and usability metrics such that a significant coefficient was revealed only between the 'performance' sub-scale of NASA-TLX questionnaire and aesthetics of the 1-back test *(r=*-0.52*, p=* 0.015*)*. The rest of the usability and cognitive load metrics did not correlate significantly with each other. Hence, the results of this section did not inform much about the relationship between usability and cognitive load metrics.

Table (IV) Correlation Analysis for 2-CRT

Test	Usability Metrics		
Measure	*Aesthetics*	*Functionally*	*Information*
Mean RT	0.00	-0.11	0.21
Correct responses	0.45*	0.63**	0.38
Longest streak	0.52*	0.63**	0.53*

*p<0.05
**p<0.01

Table (V) Correlation Analysis for 1-back test

Test	Usability Metrics		
Measure	*Aesthetics*	*Functionally*	*Information*
Mean RT	-0.28	-0.66**	-0.43
Correct responses	0.20	0.73***	0.42
Longest streak	0.30	0.75***	0.48*

*p<0.05
**p<0.01
***p<0.001

D. Clusters of Perceived Human Factors based on Participants' Demographics

Figure 5 and Figure 6 represent the clusters of participants' usability ratings in 2-CRT and 1-back tests split on the basis of their gender. It can be observed from both figures that female and male participants perceived the usability metrics completely differently from each other. Female participants rated aesthetics and information higher than functionality. On the other hand, male participants valued functionality higher than information and aesthetics of the apps.

We also illustrated clusters of participants' perceived usability on the basis of their education level to explore how the participants from three education levels perceived the usability metrics after they took 2-CRT and 1-back tests. Figure 8 and Figure 9 show that the participants at Ph.D. level were more strict in rating usability of both tests. Another information inferred from these figures is that participants who were studying in a Bachelor or Master program were inconsistent in rating the usability metrics of the 2-CRT test in contrast to their consistent rating scores during the 1-back test.

Clusters of the participants' work or study industry can be seen in Figure 11 and Figure 12, showing that those whose industries were education (N=1) and construction (N=1) tended to rate the usability metrics lower than the others. In contrast, participants who belong to the water engineering industry (N=2) valued usability of the apps higher than the others.

Participants' cognitive load measures split by their gender is illustrated in Figure 7, which shows that perceived cognitive load of the male and female participants are completely dif-

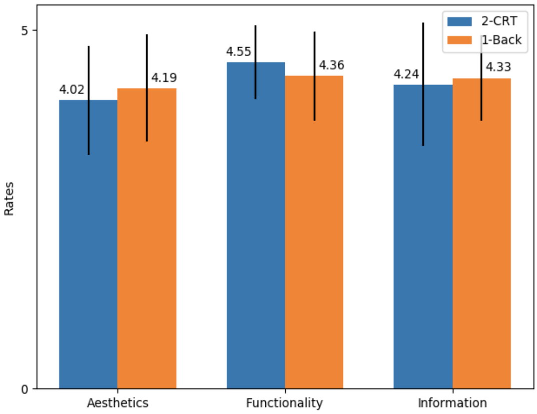


Figure (3) Usability ratings by our study participants presented separately for each cognitive test

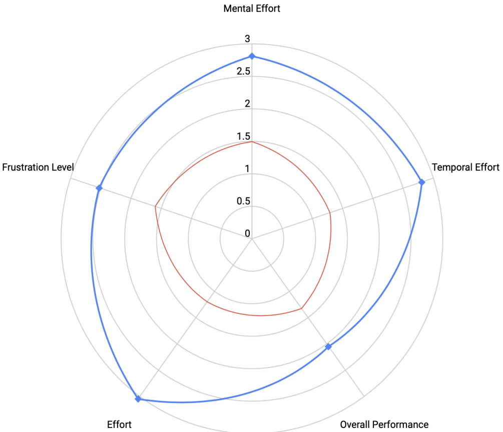


Figure (4) NASA-TLX - Sub-scales were rated by 5-point Likert scale

ferent from each other. Figure 10 represents how participants from various educational levels rated their cognitive load. As can be seen, the average of perceived frustration and performance were higher in participants at a Master's level compared to Bachelor's and Ph.D.'s level. Participants at the Bachelor level felt that their effort was high while those educating at a Ph.D. and Master program felt the opposite. Individuals at the Ph.D. level perceived higher temporal and mental demand in contrast to the participants' at the Master's level. We noticed that Master's degree participants rated their performance lower than the Bachelor and PhD level participants while Master's level scores in the 1-back test were actually higher than the Ph.D.s and a bit lower than the Bachelor's level participants.

Figure 13 shows that perceived mental effort of the participants in the computer (N=4) and design (N=4) industries were higher while the individuals who worked or studied in industries including construction, energy, food, and education perceived lower mental effort. Frustration and effort level were rated higher in food and design industries in contrast to the participant from the energy section. Temporal demand and performance were rated higher by the participant from the education section while the person in the construction industry gave a low score to the aforementioned sub-scales of the NASA-TLX questionnaire.

V. Discussion

In this study, we showed that individuals' objective cognitive performance is correlated with some metrics of their perceived usability and cognitive load. Moreover, the patterns of similarities and dissimilarities in participants' usability and cognitive load ratings were observed from the hierarchical clusters. Previous related work used questionnaires to evaluate usability of their Web-based or mobile tools. In our study, we also used a validated questionnaire including three key metrics of perceived usability. None of the previous related work investigated the associations between usability metrics of a cognitive assessment tool and their participants' cognitive test results. Furthermore, we explored users' perceived human factors on the basis of their sociodemographics to understand users' behaviour and provide insights to future application designer.

Participants' perceived human factors were associated with their cognitive performance measures. First, the significant correlation coefficients found between the functionality of the apps and participants' accuracy (see Table IV and Table V) indicate that users' behaviour in rating the usability is related to how they performed in the tests. The positive association between the longest scoring streaks and information quality and quantity shows that those who understood the instructions of the test were better in keeping the scoring streak. Second, the results reported in Section IV-B show an association between working memory performance and some sub-scales of perceived cognitive load. Higher perceived mental and time pressure led to slower RTs in the 1-back. Moreover, there was a moderate negative correlation between participants' perceived level of frustration and time pressure and both their scores and longest streaks. Given that excessive mental load have an adverse impact on learnability [3]–[5], it can be inferred that participants' frustration and stress level negatively affected their performance in the 1-back test.

Participants who studied at three educational level rated their cognitive load differently. A discrepancy was also found between perceived performance and the actual test results of the participants, indicating that participants were not able to accurately quantify their own performance level. Thus, studies that rely on users' perceived cognitive performance using subjective methods (e.g. self-reports) should consider this discrepancy.

According to our findings in Section IV-C, only perceived performance correlated with the aesthetics of the 1-back test. It can be inferred that participants rated aesthetics of the 1-back

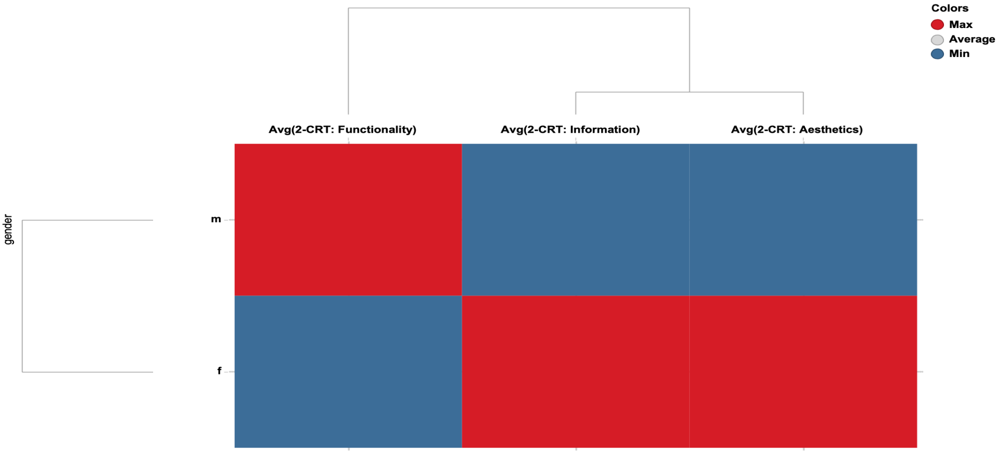


Figure (5) Clusters of participants' gender (m=male, f=female) based on their perceived usability of the two-choice reaction time test.

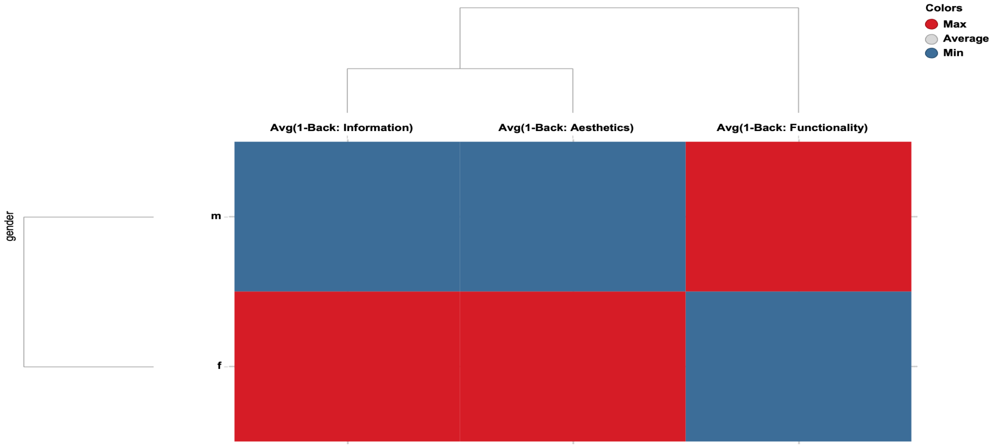


Figure (6) Clusters of participants' gender (m=male, f=female) based on their perceived usability of the 1-back App.

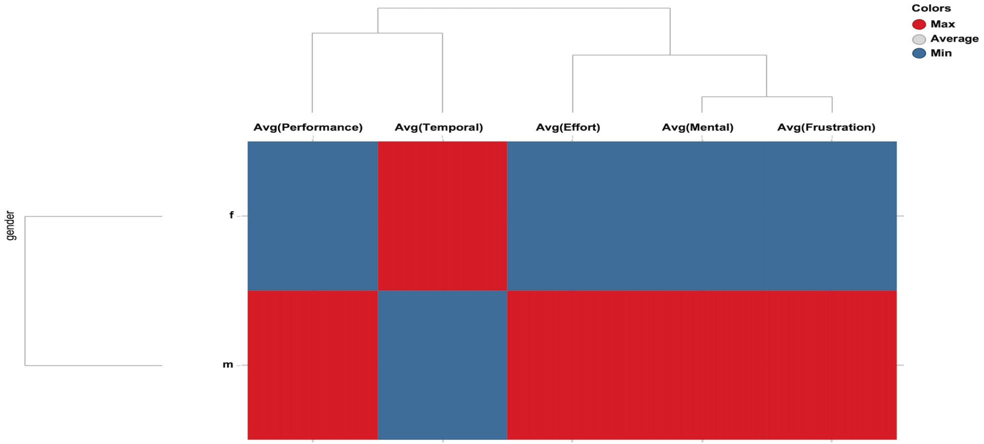


Figure (7) Clusters of participants' gender (m=male, f=female) based on their perceived cognitive load.

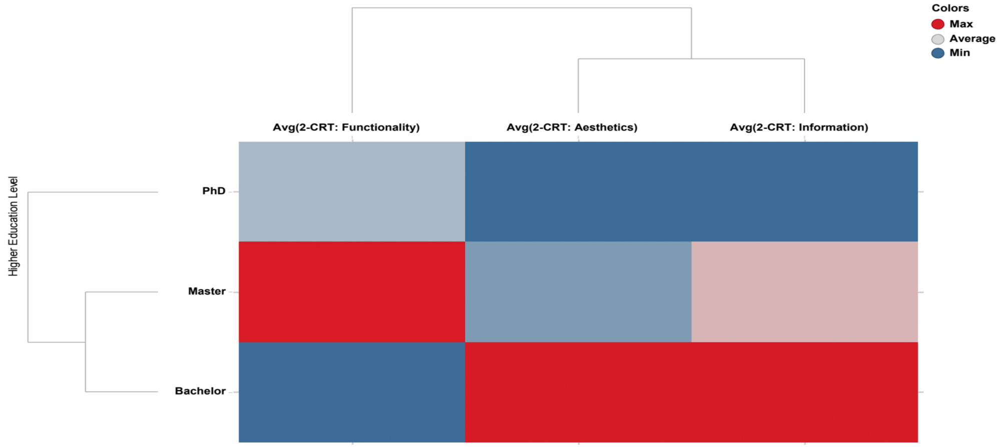


Figure (8) Clusters of participants' education based on their perceived usability of the two-choice reaction time application.

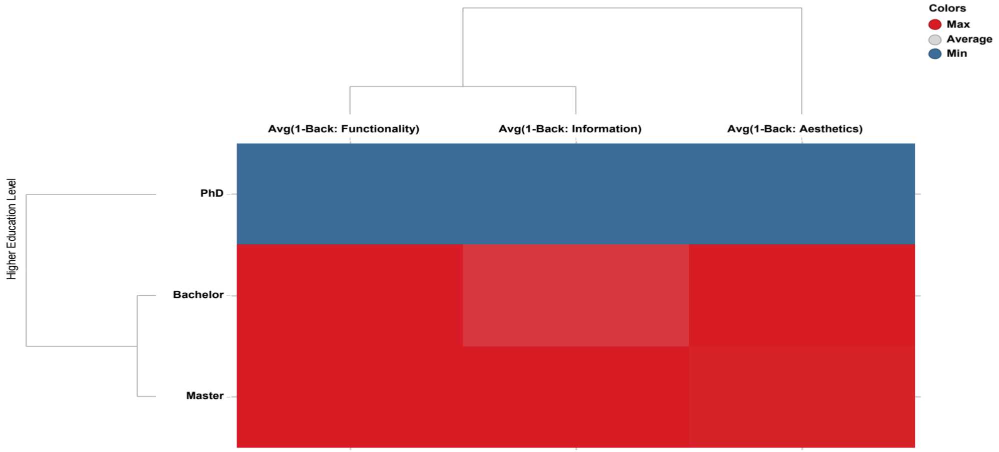


Figure (9) Clusters of participants' education based on their perceived usability of the 1-back application.

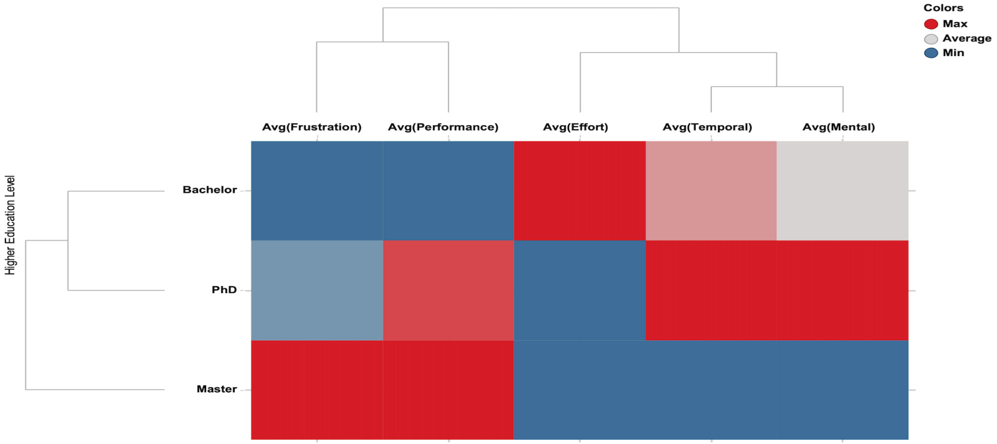


Figure (10) Clusters of participants' higher education level based on their perceived cognitive load.

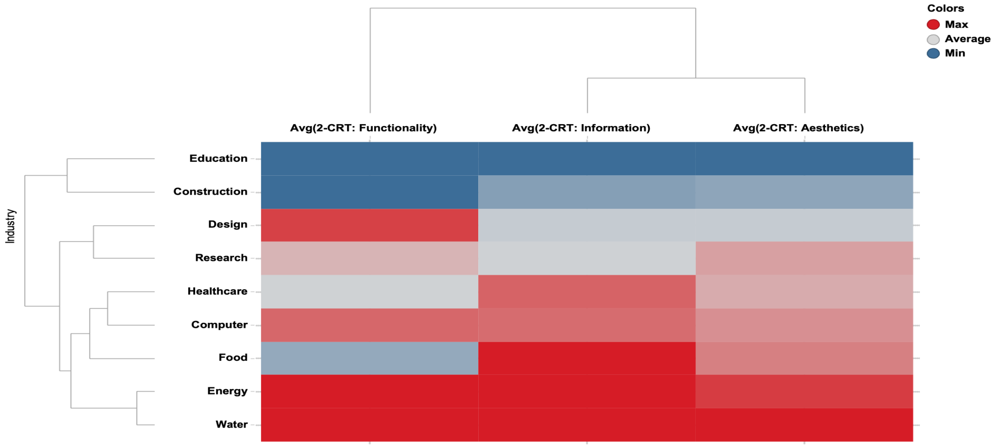


Figure (11) Clusters of participants' education or work industry based on their perceived usability of the two-choice reaction time application.

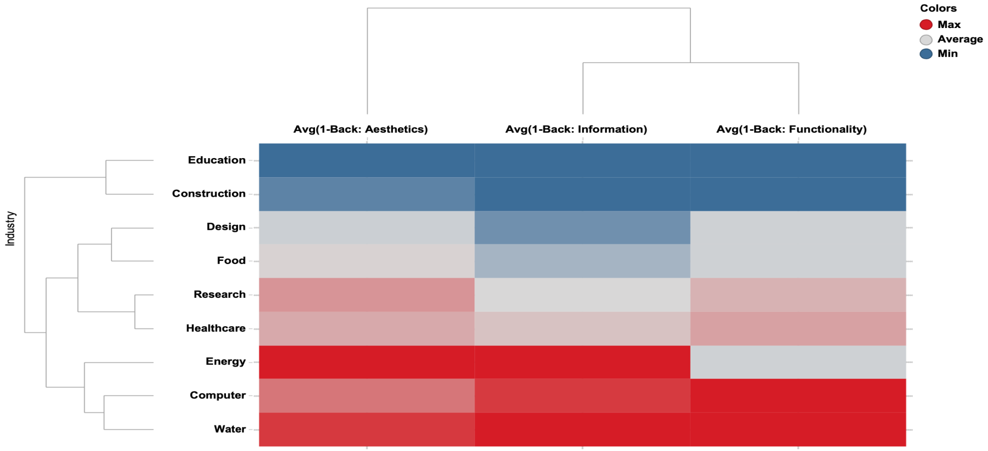


Figure (12) Clusters of participants' education or work industry based on their perceived usability of the 1-back application.

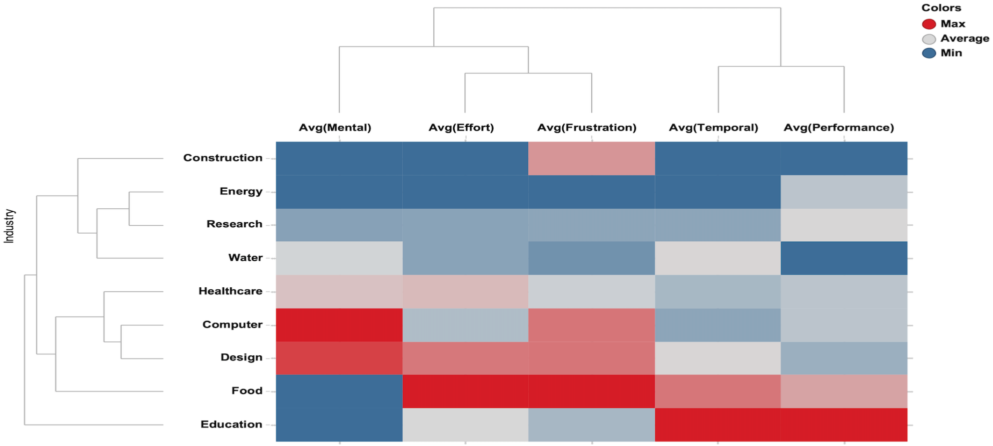


Figure (13) Clusters of participants' education or work industry based on their perceived cognitive load.

user interface inline with their perceived performance in the 1-back test while the rest of the factors did not correlate significantly with each other. A recent study showed that perceived usability and cognitive load are two independent metrics in the field of human-computer interaction [32]. As such, the correlation between perceived performance and aesthetics may not be sufficient enough to conclude any association between usability and cognitive load metrics.

Analysis performed between participants' perceived human factors and their gender and work industry also gave new insights. Female and male participants perceived the usability metrics completely differently from each other. Such a contrast shows that users' satisfaction is related to their gender. Moreover, a lack of consistency in reported usability metrics of the 2-CRT test is noticeable in design (N=4), healthcare (N=3), and food (N=1) industries. On the other hand, participants were more or less consistent in rating the usability metrics of the 1-back test. It can be inferred that user interface design of the 1-back is more acceptable than 2-CRT.

Similar to the patterns observed in Figure 5 and Figure 6, the cognitive load ratings among the male and female populations as shown in Figure 7 are completely different from each other. The perceived temporal demand in female participants was higher than the rest of the NASA-TLX sub-scales. In contrast, male participants rated their perceived temporal demand lower than the rest of the sub-scales. As temporal demand points to the pace of the app, the time limit to respond to a test stimulus may adapt to the user's gender to achieve a reliable measure of working memory. We also investigated perceived cognitive load of the participants from various industries in Section IV-D. Taken together, different patterns of perceived human factors highlight that user's satisfaction and learnability in an app are dependant on measures of sociodemographics including gender and work or study industry. In addition, adapting user interfaces to the user's characteristics may facilitate the interaction with cognitive tools to obtain reliable cognitive performance measures.

VI. Conclusion

Objective cognitive test performance measures are associated with individuals' key human factors including usability and cognitive load metric, which were evaluated subjectively. Moreover, clusters of individuals' perceived usability metrics and cognitive load sub-scales revealed patterns of similarities and dissimilarities on the basis of their sociodemographics features. Gender, education level, and work or study industry are the factors that can distinguish users of the smartwatch-based cognitive assessment tools when evaluating their perceived usability and cognitive load metrics. The findings of this study will inform the HCI and Health Informatics community about the role of human factors in designing more usable cognitive assessment technologies to achieve reliable measures of human mental health.

A. Limitation

A common issue with empirical studies to assess cognition is the challenge of recruiting a large number of participants. We have faced the same challenge in our study. The analysis performed in this study is based on a limited number of participants. We could not recruit more participants for the current study and we did not find patterns of subjective human factors based on the age of individuals.

B. Future Work

In future work, we would like to continue with larger scale studies, recruiting participants from different backgrounds and for longer period. Patients who suffer from a mental illness, for instance depression, can be the target population for future studies. Furthermore, other cognitive domains and digital cognitive assessment tools developed for other platforms can be studied to extensively explore the characteristics of their users. Finally, individuals from other work or study industries may be included in future work to be able to generalize the findings of this study. As such, a future exploration to use other clustering methods would be required since determining the correct number of clusters by the dendrograms would be difficult when using the Ward method.

Acknowledgment

This project is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 722561.

Appendix

The selected questions from the Mobile Application Rating Scale can be found here: https://doi.org/10.5281/zenodo.3364314

References

[1] ISO, *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts*, 2nd ed. Geneva, Switzerland: ISO 9241-11:2018, 2018.

[2] B. C. Zapata, J. L. Fernández-Alemán, A. Idri, and A. Toval, "Empirical studies on usability of mhealth apps: a systematic literature review," *Journal of medical systems*, vol. 39, no. 2, p. 1, 2015.

[3] S. Kalyuga, "Cognitive load theory: How many types of load does it really need?" *Educational Psychology Review*, vol. 23, no. 1, pp. 1–19, 2011.

[4] B. Xie and G. Salvendy, "Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments," *Work & stress*, vol. 14, no. 1, pp. 74–99, 2000.

[5] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and instruction*, vol. 4, no. 4, pp. 295–312, 1994.

[6] G. Lyon and N. A. Krasnegor, *Attention, memory, and executive function.* Paul H Brookes Publishing Co., 1996.

[7] C. Cognition, "Cantab mobile. luettu 24.9. 2016," 2016.

[8] J. E. Harrison, H. Barry, B. T. Baune, M. W. Best, C. R. Bowie, D. S. Cha, L. Culpepper, P. Fossati, T. L. Greer, C. Harmer *et al.*, "Stability, reliability, and validity of the thinc-it screening tool for cognitive impairment in depression: A psychometric exploration in healthy volunteers," *International journal of methods in psychiatric research*, vol. 27, no. 3, p. e1736, 2018.

[9] P. Maruff, E. Thomas, L. Cysique, B. Brew, A. Collie, P. Snyder, and R. H. Pietrzak, "Validity of the cogstate brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and aids dementia complex," *Archives of Clinical Neuropsychology*, vol. 24, no. 2, pp. 165–178, 2009.

[10] P. Hafiz, K. W. Miskowiak, L. V. Kessing, A. E. Jespersen, K. Obenhausen, L. Gulyas, K. Żukowska, and J. E. Bardram, "The internet-based cognitive assessment tool: System design and feasibility study," *JMIR formative research*, vol. 3, no. 3, p. e13898, 2019.

[11] F. Cormack, M. McCue, N. Taptiklis, C. Skirrow, E. Glazer, E. Panagopoulos, T. A. van Schaik, B. Fehnert, J. King, and J. H. Barnett, "Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: Longitudinal observational study," *JMIR mental health*, vol. 6, no. 11, p. e12814, 2019.

[12] P. Hafiz and J. E. Bardram, "Design and formative evaluation of cognitive assessment apps for wearable technologies," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 1162–1165.

[13] D. Karahoca and A. Karahoca, "Assessing effectiveness of the cognitive abilities and individual differences on e-learning portal usability evaluation," *Procedia-Social and Behavioral Sciences*, vol. 1, no. 1, pp. 368–380, 2009.

[14] C.-H. Ke, H.-M. Sun, Y.-C. Yang, and H.-M. Sun, "Effects of user and system characteristics on perceived usefulness and perceived ease of use of the web-based classroom response system." *Turkish Online Journal of Educational Technology-TOJET*, vol. 11, no. 3, pp. 128–143, 2012.

[15] R. van der Vaart, D. van Driel, K. Pronk, S. Paulussen, S. te Boekhorst, J. G. Rosmalen, and A. W. Evers, "The role of age, education, and digital health literacy in the usability of internet-based cognitive behavioral therapy for chronic pain: Mixed methods study," *JMIR formative research*, vol. 3, no. 4, p. e12883, 2019.

[16] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[17] J. Sauro, *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011.

[18] P. T. Kortum and A. Bangor, "Usability ratings for everyday products measured with the system usability scale," *International Journal of Human-Computer Interaction*, vol. 29, no. 2, pp. 67–76, 2013.

[19] P. Kortum and S. C. Peres, "Evaluation of home health care devices: Remote usability assessment," *JMIR human factors*, vol. 2, no. 1, p. e10, 2015.

[20] A. Bangor, P. Kortum, and J. Miller, "The system usability scale (sus): An empirical evaluation," *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[21] P. Kortum and F. L. Oswald, "The impact of personality on the subjective assessment of usability," *International Journal of Human–Computer Interaction*, vol. 34, no. 2, pp. 177–186, 2018.

[22] E. O. Mkpojiogu, N. L. Hashim, and R. Adamu, "Observed demographic differentials in user perceived satisfaction on the usability of mobile banking applications," 2016.

[23] F. C. Donders, "On the speed of mental processes," *Acta psychologica*, vol. 30, pp. 412–431, 1969.

[24] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." *Journal of experimental psychology*, vol. 55, no. 4, p. 352, 1958.

[25] S. R. Stoyanov, L. Hides, D. J. Kavanagh, O. Zelenko, D. Tjondronegoro, and M. Mani, "Mobile app rating scale: a new tool for assessing the quality of health mobile apps," *JMIR mHealth and uHealth*, vol. 3, no. 1, p. e27, 2015.

[26] T. Kosch, M. Hassib, P. W. Woźniak, D. Buschek, and F. Alt, "Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[27] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 835–844.

[28] B. Pfleging, D. K. Fekety, A. Schmidt, and A. L. Kun, "A model relating pupil diameter to mental workload and lighting conditions," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5776–5788.

[29] B. Mehler, B. Reimer, J. F. Coughlin, and J. A. Dusek, "Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers," *Transportation Research Record*, vol. 2138, no. 1, pp. 6–12, 2009.

[30] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[31] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[32] L. Longo, "Subjective usability, mental workload assessments and their impact on objective human performance," in *IFIP Conference on Human-Computer Interaction*. Springer, 2017, pp. 202–223.

10.4 Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls

Authors: **Pegah Hafiz**, Kamilla W Miskowiak, Lars V Kessing, Alban Maxhuni, Jakob E Bardram

Manuscript submitted to ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)

Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls

AUTHOR(S): BLIND FOR REVIEW

Mobile cognitive tests have been emerged to first, bring the assessments outside the clinics and second, frequently measure individuals' cognitive performance in their free-living environment. Patients with Bipolar Disorder (BD) suffer from cognitive impairments and poor sleep quality negatively affects their cognitive performance. Wearables are capable of unobtrusively collecting multivariate data including activity and sleep features. In this study, we analyzed daily attention, working memory, and executive functions of patients with BD and healthy controls by using a smartwatch-based tool to 1) investigate its concurrent validity and feasibility and 2) identify digital phenotypes of mental health using daily cognitive and mobile sensor data. Our findings demonstrated that the smartwatch-based tool is feasible with valid measures for *in-the-wild* cognitive assessments. Analysis showed that the patients responded slower than the healthy controls during the attention task, which may indicate lower alertness of this group. Furthermore, sleep duration correlated positively with participants' working memory performance the next day. Supervised learning models were applied to the daily cognitive and mobile features to predict individuals' mental health diagnosis. Of the models, Extreme Gradient Boosting (XGBoost) outperformed the rest of the models. In addition, feature analysis of the XGBoost ranked time in bed, daily step counts, number of missed stimuli, and measures of executive functions as the digital phenotypes of mental health diagnosis.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; **Mobile devices**; **Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: cognition, wearable technology, mobile sensing, mental health, bipolar disorder, digital phenotype

ACM Reference Format:
Author(s): Blind for review. 2018. Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls. In *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Cognitive functioning of individuals' attention, memory, and executive skills characterize the quality of their daily tasks. The common practice in psychiatry is to assess patients' cognitive functioning using neuropsychological tests. Such experiments are run in a controlled environment and at a certain time of a day suitable for the clinician and patient. However, fixed environment and context for taking cognitive tests may negatively impact the validity and reliability of the test results [2] since human cognition fluctuates during the day [49, 61]. In particular, patients with Bipolar Disorder (BD) (mania and depression) suffer from cognitive impairment even during their period of symptom remission [7, 54]. However, constraints on time and resources hinder continuous and frequent monitoring of patients' cognitive functioning. Therefore, novel computing technologies are essential to obtain cognitive performance measures over time.

Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., Vol. 0, Issue 0

https://doi.org/10.1145/1122445.1122456

A few smartphone-based tools have been proposed for measuring individuals' cognitive functioning outside the clinic [8, 25, 42, 55]. Although these tools have contributed to mobile assessments, the use of wearables propose two advantages over smartphones. First, wearables can collect reliable data on physical activity (e.g., step count) and physiological data (e.g., heart rate and sleep). Second, wearables are devices that people carry continuously and wear them during most activities such as walking and running. Recently, smartwatch-based tools have been developed to frequently assess cognitive functions [12, 20]. Taken together, wearables provide an opportunity in conducting *in-the-wild* studies for collecting multivariate sensor data in conjunction with cognitive test performance measures. It is, however, essential to evaluate the feasibility of using such tools and their concurrent validity compared with the gold-standard neuropsychological tests.

Active and passive data collected via mobile devices can assist in identifying digital phenotype of human mental health [41]. Wearables are capable of unobtrusively collecting various data types. For instance, daily step counts, physical activities, and sleep duration per cycle are calculated by Fitbit trackers. There is evidence that sleep quality affects individuals' cognitive performance during a day [19, 37]. Particularly, patients with BD can potentially suffer from the negative consequences of their frequent poor sleep quality [5]. So far, digital behavioural phenotypes of individuals' mental health have been investigated (for example, [11, 17, 18, 47, 48, 57, 58]). Yet, we do not know what cognitive, behavioural, and physiological features play a significant role in classifying individuals' mental health diagnosis.

In this study, first, we show the concurrent validity and feasibility of an *in-the-wild* cognitive assessment tool developed for Fitbit Ionic smartwatches. Then, we collect daily cognitive performance measures as well as activity and sleep features using the smartwatch to 1) investigate the impact of sleep on the next-day cognitive performance measures and 2) identify digital phenotypes of individuals' mental health diagnosis.

2 RELATED WORK

A number of studies have shown the feasibility of mobile cognitive assessments by using Personal Digital Assistants (PDAs), cellphones, or smartphones [39]. Table 1 provides an overview of studies that has used smartphone or smartwatch technology for cognitive testing. These studies have all adopted the Ecological Momentary Assessment (EMA) [52] or Experience Sampling Method (ESM) [32] methodology (which are often mentioned interchangeably). Prior related work has been focusing on collecting self-reports on mood [12, 13, 24, 56], sleep [1, 13, 16, 24], activity [13], location [13, 56], and alertness [1, 16]. No effect of sleep quality, mood, location, or activity stress was found on the cognitive test results in [13]. An 'in-the-wild' study investigated individuals' alertness and showed the effectiveness of using mobile cognitive tasks in detecting circadian variations [16]. However, self-reports on sleep duration and quality did not have an impact on the cognitive test measures. On the other hand, a similar study reported a negative impact of poor sleep on individuals' alertness [1].

Previous studies mostly collected self-reported, subjective measures of sleep and behavioural features. In this study, we collect objective sleep data using Fitbit smartwatches, which has shown acceptable performance in differentiating sleep and wake cycles [21] and in estimating sleep stage accuracy [22]. Activity features (e.g., step count) are also collected passively using the Fitbit smartwatches. Of the recent studies performed to assess *in-the-wild* cognition, two measured objective attention [1, 16], one evaluated working memory [12], and one assessed working memory and psychomotor speed [13]. We extend this work by measuring three key cognitive domains namely attention, working memory, and executive functions. Our work reports the findings of a clinical *in-the-wild* feasibility study conducted with healthy controls and patients with BD, and achieves the following contributions:

- Demonstrating the concurrent validity of the smartwatch-based tool in assessing individuals' cognitive functioning.

Table 1. Overview of the studies conducted with mobile devices for cognitive assessments.

Study	Device	Participants	Sampling	Duration	Cognitive tasks
Timmers et al. [56]	Smartphone	Young adults (N=26)	4 times daily	1 day	Letter span
Sliwinski et al. [50]	Smartphone	Adults (N=219)	5 times daily	14 days	Symbol search, Dot memory
Abdullah et al. [1]	Smartphone	Students (N=40)	2 times daily	40 days	Psychomotor vigilance test
Dingler et al. [16]	Smartphone	Students (N=12)	1–6 times daily	2–13 days	Psychomotor vigilance, Go No-Go, Multiple object tracking
Daniels et al. [13]	Smartphone	Healthy adults (N=49)	8 times daily	6 days	Visuospatial working memory, Digit symbol substitution
Hung et al. [24]	Smartphone	Depression (N=54)	once per week	8 weeks	Stroop, TMT part B
Cormack et al. [12]	Smartwatch	Depression (N=30)	3 times daily	6 weeks	N-back

- Showing the feasibility of a smartwatch-based tool for continuous, daily, *in-the-wild* administration of cognitive assessment tests.
- Investigating the relationship between sleep duration and individuals' cognitive functioning the following day.
- Identifying digital phenotypes of human mental health using daily cognitive tests combined with mobile and wearable sensor data.

3 METHODOLOGY

The methodology of our study is adapted from the EMA/ESM approach and aims at collecting active cognitive performance measures and passive mobile sensor data. This study was exempted for ethical approval by the [Blind for review] ethics committee [file: Blind for review]. All participants were informed about the types of data collected during the study and signed an informed consent before enrolled in the study (see also Figure 1). We used a cognitive assessment tool developed for Fitbit Ionic smartwatches [20] that collects daily cognitive performance measures as well as behavioural, contextual, and physiological data. The cognitive tests of this tool are choice reaction time to measure attention, N-back to evaluate working memory, and Stroop color-word test to assess executive functions. Each cognitive test of this tool is a standalone smartwatch-based app. The snapshots and description of the smartwatch-based cognitive tests are presented in Appendix A.

Table 2 shows the features collected for this study as well as their associated characteristics. As can be seen, the performance measures of the cognitive tests were the number of correct responses and response times (RTs) to the test stimuli. The feature 'missed stimuli' in Table 2 refers to the number of stimuli in the test sessions to which the participant did not respond during a time limit (2500 ms). The hit rates and false alarm rates were also calculated for the N-back test. Hit rate is the number of times that the user correctly identified a match in the N-back divided by the total matches in the sequence. False alarm rate refers to the number of times that the user responded as a match while there was no match in the sequence. The median RT was calculated for the choice

Table 2. Features collected throughout the study

#	Name	Category	Type	Features
1	Choice reaction time	Cognitive	Active	Median response time, correct responses, missed stimuli
2	N-back	Cognitive	Active	Mean response time, correct responses, hit rate, false alarm rate, missed stimuli
3	Stroop	Cognitive	Active	Mean response time, correct responses, missed stimuli
4	GPS	Contextual	Passive	Latitudes and longitudes to detect indoor and outdoor environments
5	Time of the day	Contextual	Passive	Time extracted from cognitive test logs
6	Physical Activity	Behavioural	Passive	Step counts, Minutes Sedentary, Minutes Lightly Active, Minutes Fairly Active, Minutes Very Active, Activity Calories
7	Sleep	Physiological	Passive	Minutes Asleep, Minutes Awake, Number of Awakenings, Time in Bed, Minutes REM Sleep, Minutes Light Sleep, Minutes Deep Sleep

reaction time test since this test has two stimuli, a left- and right-hand arrow, and the participants were required to select the correct direction of the arrows as fast as possible. As for the N-back and Stroop tests, the mean RT was calculated .

To establish the validity of the tool, we applied the method of ‘concurrent validity’, which is used to evaluate the measures of a novel tool against the current practice [35, 40]. Due to the frequent fluctuations in human cognition, the validity of the smartwatch-based tests was assessed immediately after the neuropsychological test sessions held at the clinic. Therefore, relevant cognitive domains of the neuropsychological tests were selected by psychiatrists to evaluate concurrent validity of the smartwatch-based cognitive tests. The tests administered during the follow-up visits included Trail Making Test (TMT) part A and B [44], Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) coding and digit span [15], Wechsler Adult Intelligence Scale Letter-Number Sequencing (WAIS LNS) [28], and verbal fluency [59]. Z-transformation of the raw scores were calculated per neuropsychological test as provided below:

(1) **Composite attention and processing speed scores**: Average of z-transformations of the scores in the RBANS coding and digit span and the TMT part A.
(2) **Composite executive functions scores**: Average of z-transformations of the scores in the verbal fluency and TMT part B.
(3) **Composite working memory scores**: Z-transformations of the scores in the WAIS LNS.
(4) **Global cognitive composite scores**: Average of z-transformations of the composite scores calculated for the attention and processing speed (1) and working memory (3).

The scores of the participants with the smartwatch-based cognitive tests were averaged and their z-transformed scores were used to compare against their global composite scores.

3.1 Study Procedure

Prior to the clinical feasibility study, we conducted a 1-week pilot study with two adults without any previous mental illness and one patient with BD in order to test data collection as well as refining the study procedure according to the participants' feedback. Then recruitment of participants was commenced at [Blind for review location] for the *in-the-wild* clinical feasibility study. The study included both healthy controls and patients with BD, who were in remission when they came to the clinic for a follow-up visit. Participants were reimbursed with an amount equivalent to 1.5 USD per cognitive test and were additionally reimbursed with an amount equivalent to 3 USD per night in case they wore the smartwatch during their sleeping time.

The study had three stages per participant as illustrated in Figure 1. Participants first underwent neuropsychological testing at the clinic. Upon finishing their follow-up visit, the study leader explained the goals and requirements to them and asked for voluntarily participation in the study. If a participant volunteered, detailed information on the study was given to them both in writing and orally, and a consent form was signed. Each participant performed the cognitive tests on the smartwatch immediately after finishing the neuropsychological tests at the clinic. Followed by that, a Fitbit Ionic smartwatch was given to the participant and s/he was instructed to perform the cognitive tests for seven days. Activity and sleep data were passively collected using the Fitbit Application Programming Interface (API).

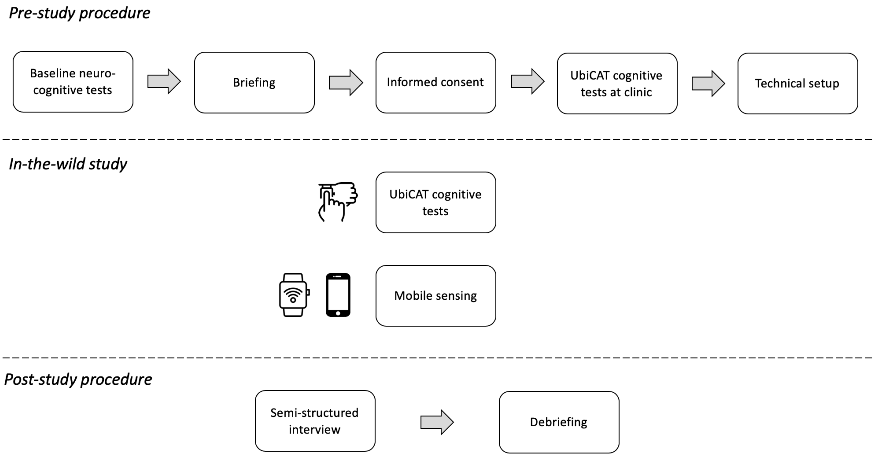


Fig. 1. Study procedure performed in three phases per participant.

Three alarms were set on the smartwatches; morning, afternoon, and evening. The hours set for the participant was not fixed for all of them due to different working or studying schedules. The study leader asked them to take the cognitive tests one by one when they received an alarm during their earliest suitable time and emphasized on taking the tests in both indoor and outdoor environments. The stimulus of each smartwatch-based cognitive test was generated randomly. Each of the N-back sessions were either 1-back, 2-back, or 3-back such that none of the two consecutive sessions were the same. For instance, if a participant finished a 1-back task, the next time it was either a 2-back or 3-back task. While participants took cognitive tests on their smartwatch, their

Global Positioning System (GPS) data was collected through the Fitbit API to detect their location. One read per second was used for the GPS. The Fitbit mobile app was installed on the participant's own smartphone since Fitbit smartwatches collect physical activity data in conjunction with their mobile app. Figure 2 illustrates a participant standing while taking a cognitive test with the smartwatch-based tool.

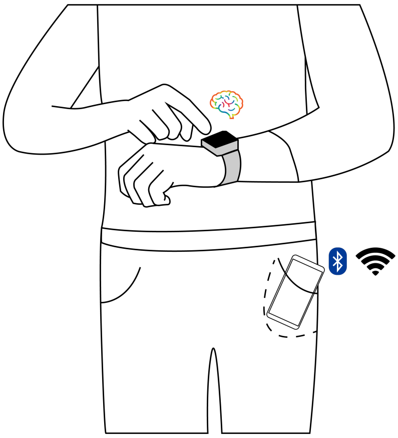


Fig. 2. A sketch of a participant taking a cognitive test with the smartwatch tool while standing. The participant's smartphone concurrently collects GPS data.

Participants were asked to return the smartwatch upon finishing the seven-day experiment. A short semi-structured interview was held with them to explore possible issues and how comfortable or uncomfortable it was to take the tests indoor or outdoor. In addition, the contexts in which they took the tests as well as their positions were inquired. It should be noted that we were not allowed to record these interview. Thus, the study leader took notes on the participants' responses during the interviews. Finally, participants were debriefed about the further analysis of the data in the study.

3.2 Data Collection and Pre-processing

The cognitive tests results were stored locally on the smartwatch. When the smartwatche was handed back at the end of the study, the test logs were extracted and transferred to spreadsheets. Activity and sleep data were stored in the Fitbit server. We screened our dataset for missing values and outliers in the cognitive, contextual, behavioural, and physiological features. As such, the RTs to the tests stimuli below 200 ms were removed since participants might have accidentally tapped on the watch screen. One participant did not wear the smartwatch during sleep and another participant only took one cognitive test daily and took off the watch during sleeping hours in most of the nights. The samples of this participant was excluded from data analysis.

The relative RTs of the participants were calculated to obtain the degree to which their alertness increased or decreased. The positive and negative values of the relative RTs shows an increase or decrease in their own alertness, respectively. Given that $MRT_{s,p}$ is the median RT of participant p in session s of the choice reaction time test, we calculated $MMRT_{s,p}$ as the mean of $MRT_{s,p}$ to obtain the $RRT_{s,p}$ (relative RT) as shown in eq. (1) (taken from [1]).

$$RRT_{s,p} = (1 - MRT_{s,p}/MMRT_{s,p}) * 100 \tag{1}$$

Two datasets were prepared for analysis of this paper. A dataset was used for analyzing the impact of sleep on the next-day cognitive performance in Section 4.3. To create this dataset, first, the accuracy of the participants were averaged separately per test for each day. Then, the participant's sleep data during one night before was added to the daily cognitive observations. The second dataset was prepared by adding participants' corresponding daily activity features to the first dataset for the purpose of digital phenotyping and training supervised models as reported in Section 4.4. This combined dataset is referred as *daily cognitive and mobile data* in the rest of the paper.

3.3 Data Analysis

Several methods were used to calculate the results of this study. Pearson correlation was used for evaluating concurrent validity of the smartwatch-based tool and the association between sleep duration and cognition. T-test was performed to compare the healthy and patient groups with each other. Analysis of Variance (ANOVA) was applied to assess feasibility of the smartwatch-based tool by measuring the impact of environment on participants' cognitive performance measures. Principal Component Analysis (PCA) [63] was performed using the *factoextra* package [27] to visualize the clusters of healthy and patients as well as the contribution of the dataset features to the principal components. The rest of the figures in this paper were created using the *ggplot2* [62] and *ggstatplot* [43] packages in R studio.

Supervised predictive models were used to classify healthy controls and patients with BD. Random Forest (RF) [34], XGBoost [10], Support Vector Machines (SVM) (radial kernel) [51], and K-Nearest Neighbour (KNN) [3] were used to build the predictive models in *caret* package [31]. Each model was trained and tested using five-fold cross validation. The label assigned to the observations was their mental health diagnosis (healthy or bipolar). It should be noted that the positive class for the predictive models was the patient's class.

4 RESULTS

We initially recruited 10 healthy controls and 8 patients with BD. Of the participants, 9 healthy controls and 6 patients completed their seven-day experiment ($N = 15$). Table 3 reports gender, age, education years of the participants, Hamilton Depression Rating Scale (HAMD) and Young Mania Rating Scale (YMRS) clinical ratings, and verbal intelligence quotients for each group. In total, we collected the following number of observations per smartwatch-based cognitive test: 318 for the choice reaction time, 294 for the N-back, and 309 for the Stroop test. Participants used the smartwatch between 6 and 18 days (Mean:8.6, SD:2.80). It should be noted that the participant who took the tests for 18 days kept the device longer than the rest of the participants due to a problem in handing back the device.

4.1 Validity and Feasibility of the In-the-wild Tool

The participants' smartwatch-based test results were used to investigate the concurrent validity of the tool. Table 4 shows the correlation coefficients applied between the neuropsychological tests and the smartwatch-based tests per cognitive domain as well as the global cognition. We found a strong, significant correlation between the average scores obtained from the smartwatch-based and neuropsychological tests *(r=0.77)* indicating adequate concurrent validity of the smartwatch-based tool. The cognitive domains also correlated significantly *(r=0.58-0.64)*. It should be noted that correlation analysis for attention and processing speed was performed between two cognitive test scores of the smartwatch-based tool namely choice reaction time and Stroop's score of congruent stimuli. Although the analysis for the Stroop's congruent scores did not reveal a significant correlation coefficient,

Table 3. Characteristics of study participants reported separately for patients and controls.

Characteristic	**Measure**	**Statistics**	
		HC	**BP**
Gender	Female	5	5
	Male	4	1
Age	Mean±SD	34±13	32±6
Years of education	Mean±SD	16±1.8	15±2.04
HAMD	Mean±SD	1.1±1.3	5.2±3.5
YMRS	Mean±SD	0.7±2	2.3±3.2
Verbal Intelligence Quotient	Mean±SD	115±5	108±3

the choice reaction time test showed a significant *p*-value indicating validity of this test for measuring attention and processing speed.

Table 4. Pearson correlation analysis between neuropsychological tests and smartwatch-based test scores.

Cognitive Function	**Neuropsychological Test**	**Smartwatch Test**	**Pearson's r**	**p**
Executive functions	Verbal fluency and TMT-B	Stroop's score to incongruent stimuli	0.58	**0.024**
Working memory	WAIS LNS	N-Back	0.63	**0.011**
Attention and processing speed	TMT-A and RBANS	Choice reaction time	0.64	**0.010**
		Stroop's score to congruent stimuli	-0.11	0.686
Global cognition	Working memory and attention	Composite scores	0.77	<**0.001**

Feasibility of our study instrument was examined to demonstrate the viability of smartwatches in assessing *in-the-wild* cognitive functioning considering the impact of indoor and outdoor places on participants' cognitive performance measures together with the interviews conducted with them. Of all participants, five patients and seven healthy individuals took the smartwatch-based tests both in indoor and outdoor environments according to their GPS data. Table 5 shows that their performance measures in all of the tests were statistically not different in the indoor and outdoor places demonstrating the feasibility of the smartwatch-based tests in individuals' free-living context. Seven participants could allocate time for the post-study interview. None of them mentioned any issue except for one participant: "I felt uncomfortable when I was together with people". Their position while taking the cognitive tests were also investigated; four participants reported *sitting* as their most common position while two reported that they sometimes took the tests while *traveling* (e.g., on the bus). Two participants took the tests while *walking*. One patient and one healthy participant reported that they were motivated by using the smartwatch, stating that: "It motivated me to track my data and I checked my activities all the time", and "The watch motivated me to walk more".

Table 5. Analysis of variance applied to examine the impact of indoor and outdoor places on the tests measures

Smartwatch-based Test	**Observations (Nr.)**	**Performance Measure**	**Mean Square**	**F**	**p**
Choice-reaction time	249	Median response times	4505.44	0.45	0.501
		Num. of correct responses	0.41	0.35	0.553
N-Back	217	Mean response times	15478.79	0.34	0.560
		Num. of correct responses	9.38	0.09	0.759
Stroop	227	Mean response times	14294.25	0.30	0.583
		Num. of correct responses	2.50	0.57	0.452

4.2 Alertness Per Hour

Individuals' alertness fluctuates during the day [49] and the choice reaction time paradigm tests are typically used for measuring individuals' alertness. Hence, we used the choice reaction time test of the smartwatch-based tool to measure and compare the alertness of the patients and healthy controls. The records of one healthy participant in the choice reaction time test were removed since this participant had some issues with the touch sensitivity of the smartwatch screen. Figure 3 shows the median RTs per group and shows that the median RTs of the healthy group were statically lower than the patients *(t(230.30)=5.24, p<0.001)* indicating better alertness of the healthy controls.

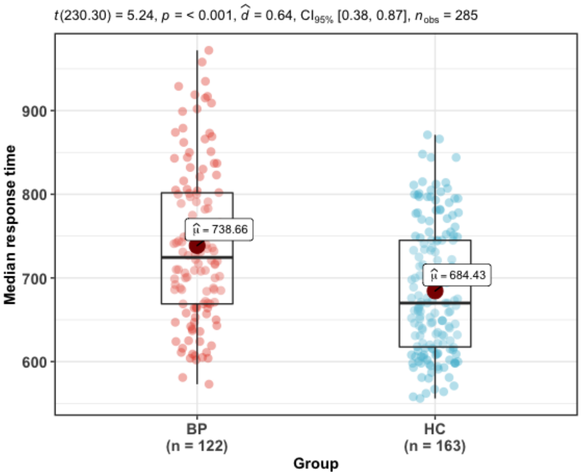


Fig. 3. Median response times of the healthy and patient groups in the choice reaction time test.

Figure 4 represents the participants' median RTs for each hour that they performed the tests. Both groups did the choice reaction time test mostly at 9AM (patient:16, healthy:7), 1PM (patient:14, healthy:17), 2PM (patient:18, healthy:16), and 6PM (patient:20, healthy:12). Independent samples t-test revealed that patients responded significantly slower at 2PM *(t(32)=3.52, p<0.001)* and 6PM *(t(30)=4.10, p<0.001)*. The relative RTs were calculated per participant similar to the approach used in previous work [1, 29, 60]. Figure 5 shows the percentage to which individuals' alertness increased or decreased compared with their own median RTs in each hour, showing an almost balanced number of positive (*N*=153) and negative (*N*=132) samples. The ratio of negative samples of the patients (48%) was higher than healthy controls (45%) while the ratio of the positive samples of healthy controls (55%) was higher than the patients (52%). Thus, the drop in alertness was higher in the patients while the rise in alertness was higher in the controls.

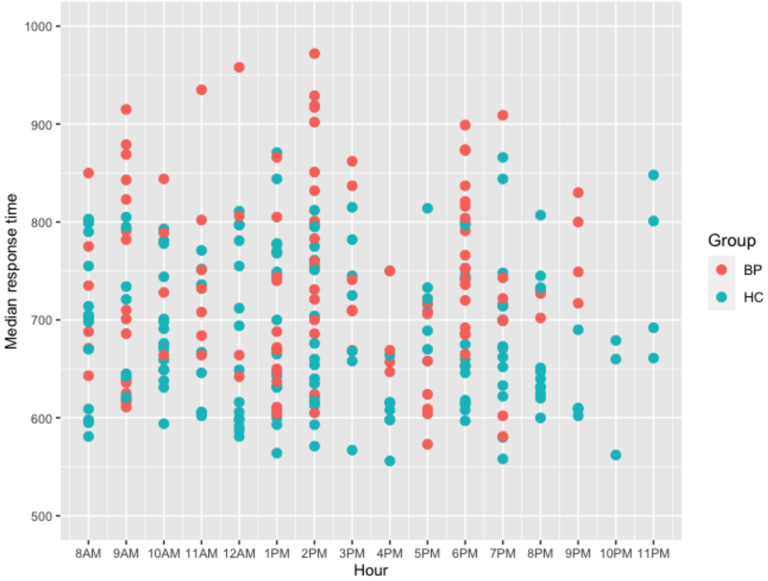


Fig. 4. Hourly representation of the median response times in the choice reaction time test.

4.3 Correlation between Sleep Duration and Cognition

Sleep duration of the participants is visualized in Figure 6 showing that patients slept more than healthy controls *(t(98.26)=3.68, p<0.001)*. Correlation analysis between sleep duration and the next-day cognitive performance measures revealed a significant coefficient in terms of the N-back hit rates as a measure of working memory performance *(r=0.26, p=0.026)*. It can be inferred that more sleeping led to higher accuracy in recalling the letters during the N-back test sessions. The rest of the cognitive test measures did not reveal any significant correlation with sleep duration. Minutes of light and deep sleep also did not correlate with the cognitive test performance measures. The difference in the participants' N-back hit rates and sleep duration are depicted in Figure 7 per group. The ratios in both plots were calculated by taking the first value (day one) as the basis to compare to the next days. While the difference in the sleep and hit rates of the two groups do not have the same pattern, there is a similar trend in their hit rates and sleep duration within groups.

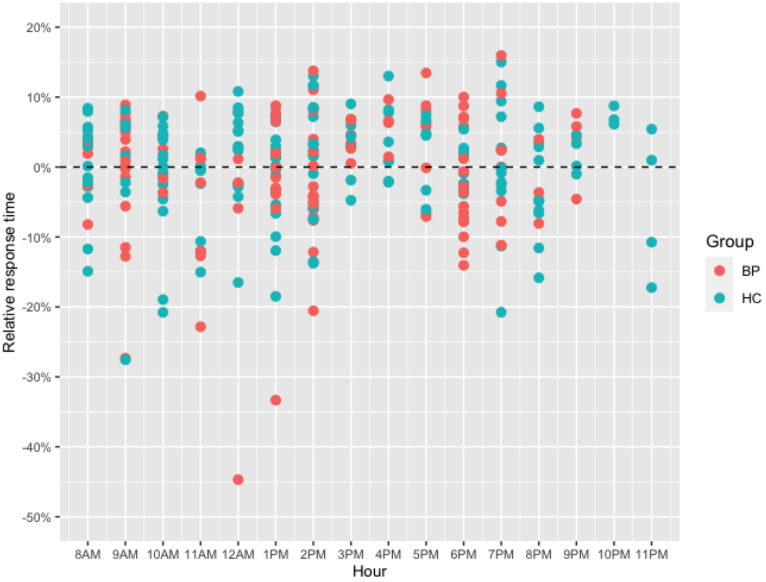


Fig. 5. Relative response times of the participants in the choice reaction time test per hour.

4.4 Daily Cognitive and Mobile data

The daily cognitive and mobile data had $N = 81$ observations. Five incomplete cases were removed from this dataset such that $N = 76$ observations remained (bipolar:40, healthy: 36). The features of this dataset were *time in bed, sleep duration, number of awakenings, minutes awake, step counts, mean RTs during the Stroop and choice reaction time tests, and average accuracy in the N-back, Stroop, and choice reaction time tests.* It should be noted that the RTs of the N-back tests were not included since the mean RTs of the tests with various difficulty levels were not comparable .

4.4.1 Statistics and Comparison. Table 6 reports the descriptive statistics of the total daily missed counts during the cognitive test sessions and the average accuracy of the participants in each test. The missed count is the number of times that a participant did not respond to the test stimuli throughout a test session. As such, daily missed count is the sum of missed counts calculated for the test sessions that the participant completed each day. T-test revealed that the patients had more daily missed counts compared with the healthy controls *(t(72)=3.24, p<0.001)*, indicating inability of the patients with BD in responding during the time limit of the cognitive tests. The average accuracy of the healthy controls in their daily cognitive tests with the smartwatch was higher than the patients although t-test analysis did not give significant p-values. The RTs during the Stroop test was averaged per day for each participant and t-test showed a significantly higher RTs of the patients *(t(72)=1.93, p=0.029)*. Hence, patients were slower in selecting the correct ink color of the stimuli in the Stroop test (see fig. 11c). The participants' daily step counts are shown in Figure 8a, showing that patients' mobility was significantly

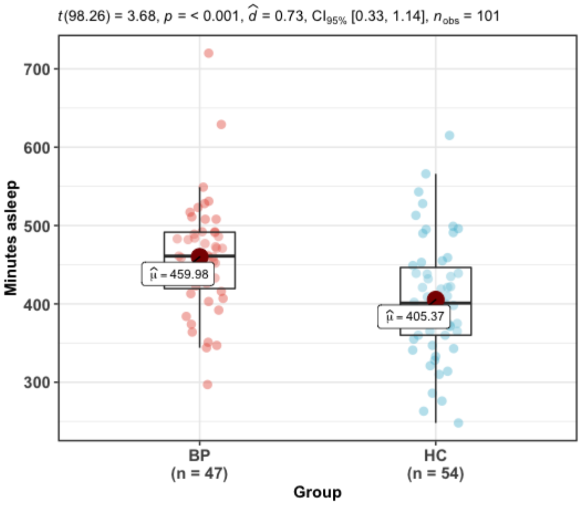


Fig. 6. Sleep duration in minutes visualized for healthy and patient groups.

Table 6. Descriptive statistics of the daily missed counts and average accuracy in the cognitive tests

	Missed Counts		CRT accuracy (%)		N-back accuracy (%)		Stroop accuracy (%)	
	BP	HC	BP	HC	BP	HC	BP	HC
Mean	3.17	1.35	98.98	99.20	90.59	92.49	96.15	96.80
SD	3.67	1.66	1.35	1.47	5.24	4.26	2.75	2.87
Minimum	0.00	0.00	95.00	92.00	78.67	83.50	88.00	87.00
Maximum	18.00	7.00	100.00	100.00	100.00	100.00	100.00	100.00

higher than the controls *(t(78.75)=2.03, p=0.046)*. Patients spend more time in bed compared with the controls *(t(78.35)=3.46, p=0.001)* as can be observed in Figure 8b.

4.4.2 Predictive Models and Feature Analysis. PCA was performed to explore the clusters and features in the dataset. The contribution of the dataset features are represented in Figure 9a. It can be seen that sleep duration, time in bed, number of awakenings, daily missed counts, and Stroop RT contributed more to the principal components 1 and 2. Figure 9b shows the clusters of healthy and patient observations as well as the overlap between the samples.

The performance evaluation results are presented in Table 7. The average accuracy of the XGBoost and KNN ($K = 9$) are very close to each other although the standard deviation is lower in the XGBoost. The average of Positive Predictive Value (PPV) in the KNN is also the highest while Negative Predictive Value (NPV) of the XGBoost is above all. The average sensitivity and specificity values of the SVM and KNN models are the highest, respectively. Overall, the XGBoost model gave the highest average Area under the Receiver Operating

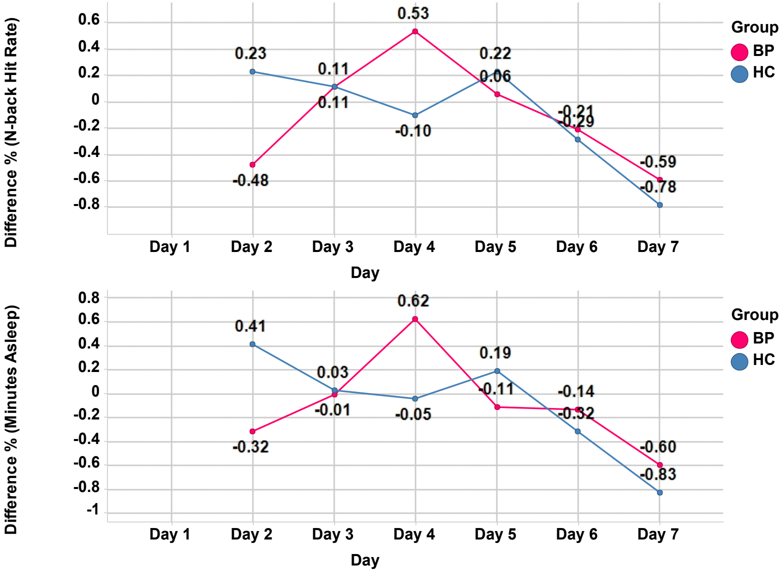


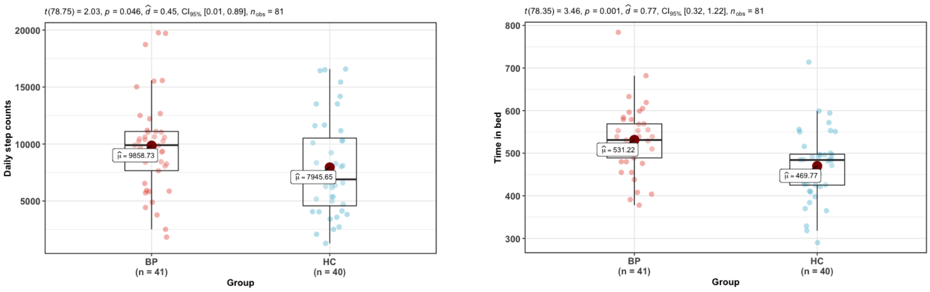
Fig. 7. Overall difference of N-back hit rates and minutes asleep per group.

Characteristic Curve (AUC). We derived the feature importance from the XGBoost trained model and represented the features versus their relative importance in Figure 10. As can be observed, *time in bed and daily step counts, total missed counts, and performance measures of the Stroop test* are the most significant digital phenotypes of participants' mental health diagnosis.

Table 7. Mean and standard deviations of the performance evaluation metrics for classification of healthy and patients.

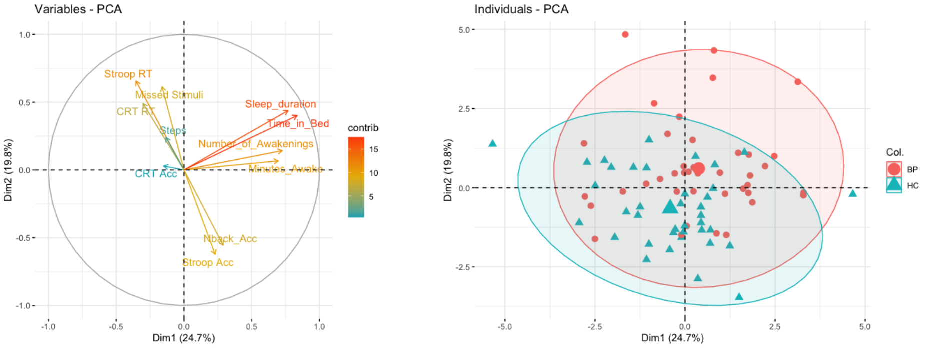
Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC (%)
XGBoost	**77.51±3.28**	76.65±2.91	78.38±4.03	78.60 ±3.83	**76.38±3.45**	**86.40±3.97**
KNN	**77.89±4.41**	73.22±10.47	**83.95±3.47**	**83.88±7.18**	71.71±10.85	80.59±1.28
RF	72.63±1.57	69.40±5.32	75.85±3.35	74.65±1.44	71.00 ±2.67	79.60±2.19
SVM	74.03±3.14	**81.53±4.03**	64.18±7.31	74.13±2.48	73.63±5.31	79.80±5.17

XGBoost: Extreme Gradient Boosting; KNN: K-Nearest Neighbour; RF: Random Forest; SVM: Support Vector Machines



(a) Daily step counts. (b) Time in bed in minutes.

Fig. 8. Daily step counts and time in bed of healthy and patient groups.



(a) Dataset features and their contributions. (b) Dataset samples and clusters.

Fig. 9. Principal Component Analysis.

5 DISCUSSION

The results from this study show that smartwatch-based cognitive assessment of cognitive functioning correlated moderately to strongly *(r=0.58-0.77)* with state-of-art neuropsychological tests, thus demonstrating adequate concurrent validity of smartwatch-based cognitive tests. Furthermore, the study shows the feasibility of *'in-the-wild'* assessment of cognitive functioning since participants' cognitive performance measures were statistically the same in both indoor and outdoor environments. To our knowledge, this study is the first that 1) demonstrated feasibility of assessing key cognitive functions of the patients with BD and healthy controls in their free-living context and 2) identified digital phenotypes of individuals' mental health diagnosis from a dataset including their daily cognitive, behavioural, and physiological features.

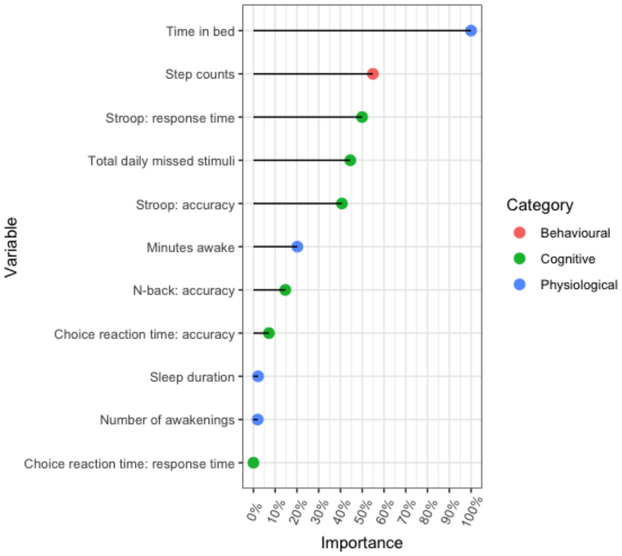


Fig. 10. Variable importance of the features ranked by the XGBoost model

5.1 Comparing Healthy and Patient Groups

Our findings showed that patients with BD responded slower compared with healthy controls during the choice reaction time test although the patients were in remission. This is in line with prior research showing that cognitive impairment remains in patients with BD despite being in remission [7, 54]. We also showed that 1) patients responses were statistically slower than the controls at 2PM and 6PM and 2) there were more drop and less rise in patients' alertness compared with the healthy controls (see Figure 5). The meaningful difference between alertness level of healthy and patient groups together with hourly-basis analysis of their alertness may inform the Ubicomp community to consider individuals' mental health diagnosis as a potential feature for managing attention-demanding tasks per hour. Prior work has established the threshold for impairment in alertness using smartphones as 500 ms in [6]. According to Figure 3, the results from this study seems to imply that the threshold of median RTs for distinguishing patients and healthy controls is 800 ms when using a smartwatch-based tool. The difference between the modality of smartphones and smartwatches justifies the difference between individuals' RTs.

We showed that the patients with BD slept significantly more than the healthy group (see Figure 6) that is inline with the findings of a study conducted with controls and patients with BD using actigraphy [38]. Figure 8b also showed that the patients stayed more time in bed compared with the healthy controls. Sleep disturbance of the patients with BD was previously demonstrated in [45] and more time in bed may inform about sleep disturbance of the patients. Mobility of the patients in terms of their step counts was higher than the controls (see Figure 8a) which might be due to the motivation caused by using the Fitbit smartwatches to walk more, which was inferred from the interviews with the participants in Section 4.1.

The daily average accuracy of the healthy group was slightly higher than the patients although the t-test did not reveal any significant p-value (see Table 6). Daily missed counts in the cognitive tests sessions were calculated per participant. Patients on average missed more stimuli than healthy participants. The higher number of daily missed count indicates an inability to respond within a time limit. While none of the previous related work investigated individuals' missed count particularly patients with BD, we showed that this objective measure has a great potential in classifying patients with BD and healthy controls.

5.2 Sleep and Cognitive Functions

One of the cognitive performance measures of the N-back test was hit rates. The higher the hit rate, the better ability of the user in keeping the letters in his/her mind and thus better working memory. We showed a significant correlation between sleep duration and the next-day working memory performance in terms of the N-back hit rates (see Section 4.3). Such finding is in line with the results of the study by Russo et al. [46] who demonstrated a negative impact of sleep disturbance on working memory for patients with BD. Moreover, cognitive load is induced by learning tasks and it directly involves working memory ability [14] while too much cognitive load adversely affects learnability [26, 53, 64]. The significant correlation coefficient between sleep duration and working memory of the participants may inform the Ubicomp community about adjusting learnability of the smartphone-based tasks according to the user's sleep duration the night before.

Similar to Dingler et al. [16], we did not find a significant impact of sleep on alertness although we used objective sleep measures. However, another study [1] did find a significant impact of sleep variation on individuals' alertness while their study aimed to systematically measure circadian rhythm during 40 days, their sample size was larger ($N = 20$) and a different target population (young college students) took part in their study.

5.3 Digital Phenotyping through Supervised Models

Various supervised learning models were applied on the daily cognitive and mobile dataset to train and test the models and select the best performing model by comparing their average AUC with each other. KNN is a simple and non-parametric classifier without any assumption about the underlying data [23] while it is susceptible to the noise in the data. However, the impact of noise is less significant when using simple classifiers like KNN rather than more complicated methods like RF and SVM [65]. SVM is suitable when classes are separable but does not perform well in case of overlapped classes. In our dataset, we observed an overlap between the samples in the clusters of healthy and patient groups (see Figure 9b). Such an overlap may justify lower values of the SVM performance metrics. The average AUC of the KNN was higher than the RF and SVM, but not as good as the XGBoost. Nevertheless, KNN performed better than the RF and SVM in predicting mental health diagnosis. XGBoost uses an ensemble of decision trees and is an enhanced algorithm of gradient boosting [10]. RF is also a tree-based method but its performance was not as good as the XGBoost. Above all, the XGBoost model gave the highest average AUC.

The ranking of the important features in the output of the XGBoost model revealed the digital phenotypes of individuals' mental health diagnosis using daily measures of their cognition, activity, and sleep features. Time in bed and step counts were the top features that were collected unobtrusively during the study. The next important features are the RTs to the Stroop test stimuli (color names) and daily missed counts of the participants. Missed count refers to individuals' ability in responding during the time limit of each test stimuli, which was 2500 ms in our study. The Stroop test displays a sequence of color names that are either congruent or incongruent (see Figure 11c) and the user should select the ink color as fast as possible without getting distracted by the meaning of the color name. As such, faster RTs in selecting the correct color shows better executive functioning.

Taken together, wearable computing technology assists in collecting frequent cognitive performance measures such that an increase in the 1) RT to the Stroop color names and 2) daily missed counts in the cognitive tests may indicate the need for a follow-up session to assess individual's mental health with a clinician. The rest of the features also were ranked as important although with comparably lower ratio. Participants' average accuracy during the N-back test slightly contributed to the variable importance of the XGBoost while the performance measures of the choice reaction time did not have a significant role in classifying patients and healthy controls.

5.4 Perspectives

The results of this study justify the feasibility of utilising objective sleep and activity data in cognition-aware systems to help in managing the demand on users' working memory the following day by, for example, reducing task load in case of poor sleep quality. Furthermore, our findings paves the way for building clinical decision support systems using wearable and mobile sensor data for timely detection of the mental disorders in particular BD. Daily observations of users' cognition and mobile data can be utilised in predicting the probability of cognitive impairments to diagnose mental disorders. Continuing this line of research will also enable researchers to include mobile and wearable sensor data in their studies to identify other digital phenotypes of cognition. One such feature is ambient noise, which can be collected via the smartphone's microphone. Possible associations between cognitive performance measures and moment-by-moment stress ratings may also provide new knowledge. Moreover, phone interaction features such as gestures and accelerometer measures can be integrated with the features we collected in our study toward a more comprehensive identification of digital phenotypes of mental health.

5.5 Limitation

This study have some limitations. First, the final sample size (number of participants) was smaller than planned due to the COVID-19 outbreak, which required recruitment to be stopped before the study ended. Nevertheless, our findings still provide a statistically significant analysis on cognitive performance measures and daily mobile data in particular related to the concurrent validity and feasibility of the smartwatch-based tool and the identification of digital phenotypes of mental health. Second, the fluctuations in the RTs over the course of the day depend on several factors including the chronotype of the individuals (for example, morningness vs, eveningness), which we did not control for this study. Third, the golden standard for sleep assessment in clinical studies are typically self-reported sleep assessments [9], activity patterns of wrist-worn actigraphy [4], or polysomnography for sleep monitoring [30, 33, 36]. Even though acceptable performance of the Fitbit device in collecting sleep data have been demonstrated [21, 22], such consumer sleep tracking devices are not medical devices and might not be as accurate.

6 CONCLUSION

This study showed that a smartwatch-based cognitive assessment tool is a valid instrument for measuring attention, working memory, and executive functions. Moreover, this tool is feasible for frequent assessment of the key cognitive functions *'in-the-wild'*, i.e. in both indoor and outdoor environments, as well as when users are taking different positions such as *sitting, standing, and walking*. We also showed the potential of wearable computing technology in identifying individuals' mental health diagnosis by collecting daily multivariate, active and passive data. Patients with BD responded slower in the attention test compared to healthy controls, indicating lower alertness level of the patients. Sleep duration correlated positively with the next-day working memory performance, which may help inform the design of cognition-aware computing system when cognitive load is managed in accordance with sleep duration. Digital phenotypes of mental health were derived from supervised models of the patients with BD and healthy controls. The time participants stayed in bed as well as their daily step counts were the most important features. Moreover, daily missed counts and response times (RTs) in the Stroop test, which measure executive function, were found to be the next significant features. We conclude that using mobile and wearable technology for ambulatory collection of individuals' physiological, behavioural and cognitive features provides the basis for assisting clinicians in continuously monitoring patients symptoms for early diagnosis and treatments.

7 ACKNOWLEDGMENTS

Blind for review.

REFERENCES

[1] Saeed Abdullah, Elizabeth L Murnane, Mark Matthews, Matthew Kay, Julie A Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 178–189.

[2] Michèle Allard, Mathilde Husky, Gwénaëlle Catheline, Amandine Pelletier, Bixente Dilharreguy, Hélène Amieva, Karine Pérès, Alexandra Foubert-Samier, Jean-François Dartigues, and Joel Swendsen. 2014. Mobile technologies in the early detection of cognitive decline. *PLoS One* 9, 12 (2014), e112197.

[3] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

[4] Sonia Ancoli-Israel, Jennifer L Martin, Terri Blackwell, Luis Buenaver, Lianqi Liu, Lisa J Meltzer, Avi Sadeh, Adam P Spira, and Daniel J Taylor. 2015. The SBSM guide to actigraphy monitoring: clinical and research applications. *Behavioral sleep medicine* 13, sup1 (2015), S4–S38.

[5] S Ancoli-Israel and Th Roth. 1999. Characteristics of insomnia in the United States: results of the 1991 National Sleep Foundation Survey. I. *Sleep* 22 (1999), S347–53.

[6] Mathias Basner, Daniel Mollicone, and David F Dinges. 2011. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta astronautica* 69, 11-12 (2011), 949–959.

[7] E Bora and A Özerdem. 2017. Meta-analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia. *Psychological medicine* 47, 16 (2017), 2753–2766.

[8] Robert M Brouillette, Heather Foil, Stephanie Fontenot, Anthony Correro, Ray Allen, Corby K Martin, Annadora J Bruce-Keller, and Jeffrey N Keller. 2013. Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. *PloS one* 8, 6 (2013), e65925.

[9] Colleen E Carney, Daniel J Buysse, Sonia Ancoli-Israel, Jack D Edinger, Andrew D Krystal, Kenneth L Lichstein, and Charles M Morin. 2012. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 35, 2 (2012), 287–302.

[10] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[11] Chul-Hyun Cho, Taek Lee, Min-Gwan Kim, Hoh Peter In, Leen Kim, and Heon-Jeong Lee. 2019. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *Journal of medical Internet research* 21, 4 (2019), e11029.

[12] Francesca Cormack, Maggie McCue, Nick Taptiklis, Caroline Skirrow, Emilie Glazer, Elli Panagopoulos, Tempest A van Schaik, Ben Fehnert, James King, and Jennifer H Barnett. 2019. Wearable Technology for High-Frequency Cognitive and Mood Assessment in Major Depressive Disorder: Longitudinal Observational Study. *JMIR Mental Health* 6, 11 (2019), e12814.

[13] NEM Daniëls, SL Bartels, SJW Verhagen, RJM Van Knippenberg, ME De Vugt, and Ph AEG Delespaul. 2020. Digital assessment of working memory and processing speed in everyday life: Feasibility, validation, and lessons-learned. *Internet Interventions* 19 (2020), 100300.

[14] Robin Deegan. 2013. Mobile Learning Application Interfaces: First Steps to a Cognitive Load Aware System. *International Association for Development of the Information Society* (2013).

[15] Faith Dickerson, John J Boronow, Cassie Stallings, Andrea E Origoni, Sara K Cole, and Robert H Yolken. 2004. Cognitive functioning in schizophrenia and bipolar disorder: comparison of performance on the Repeatable Battery for the Assessment of Neuropsychological Status. *Psychiatry research* 129, 1 (2004), 45–53.

[16] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. 2017. Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 47.

[17] Maria Faurholt-Jepsen, Jonas Busk, Helga Þórarinsdóttir, Mads Frost, Jakob Eyvind Bardram, Maj Vinberg, and Lars Vedel Kessing. 2019. Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Australian & New Zealand Journal of Psychiatry* 53, 2 (2019), 119–128.

[18] Maria Faurholt-Jepsen, Mads Frost, Maj Vinberg, Ellen Margrethe Christensen, Jakob E Bardram, and Lars Vedel Kessing. 2014. Smartphone data as objective measures of bipolar disorder symptoms. *Psychiatry research* 217, 1-2 (2014), 124–127.

[19] Marcos G Frank. 2006. The mystery of sleep function: current perspectives and future directions. *Reviews in the Neurosciences* 17, 4 (2006), 375–392.

[20] Pegah Hafiz and Jakob E Bardram. 2019. Design and formative evaluation of cognitive assessment apps for wearable technologies. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1162–1165.

[21] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. 2019. Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. *Journal of medical Internet research* 21, 11 (2019), e16273.

[22] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. 2020. Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages. *Chronobiology International* 37, 1 (2020), 47–59.

[23] D Hand, H Mannila, and P Smyth. 2001. Principles of Data Mining". The MIT Press. In *A comprehensive, highlytechnical look at the math and science behind extracting useful information from large databases*. Vol. 546.

[24] Shan Hung, Min-Shan Li, Yen-Lin Chen, Jung-Hsien Chiang, Ying-Yeh Chen, and Galen Chin-Lun Hung. 2016. Smartphone-based ecological momentary assessment for Chinese patients with depression: An exploratory study in Taiwan. *Asian journal of psychiatry* 23 (2016), 131–136.

[25] Susan Jongstra, Liselotte Willemijn Wijsman, Ricardo Cachucho, Marieke Peternella Hoevenaar-Blom, Simon Pieter Mooijaart, and Edo Richard. 2017. Cognitive testing in people at increased risk of dementia using a smartphone app: the iVitality proof-of-principle study. *JMIR mHealth and uHealth* 5, 5 (2017), e68.

[26] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review* 23, 1 (2011), 1–19.

[27] Alboukadel Kassambara and Fabian Mundt. 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses.* https://CRAN.R-project.org/package=factoextra R package version 1.0.7.

[28] Alan S Kaufman and Elizabeth O Lichtenberger. 2005. *Assessing adolescent and adult intelligence.* John Wiley & Sons.

[29] Matthew Kay, Kyle Rector, Sunny Consolvo, Ben Greenstein, Jacob O Wobbrock, Nathaniel F Watson, and Julie A Kientz. 2013. PVT-touch: adapting a reaction time test for touchscreen devices. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 248–251.

[30] Bhanu Prakash Kolla, Subir Mansukhani, and Meghna P Mansukhani. 2016. Consumer sleep tracking devices: a review of mechanisms, validity and utility. *Expert review of medical devices* 13, 5 (2016), 497–506.

[31] Max Kuhn. 2020. *caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret R package version 6.0-86.

[32] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology*. Springer, 21–34.

[33] Zilu Liang and Mario Alberto Chapa Martell. 2018. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *Journal of Healthcare Informatics Research* 2, 1-2 (2018), 152–178.

[34] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[35] Wei-Ling Lin and Grace Yao. 2014. *Concurrent Validity.* Springer Netherlands, Dordrecht, 1184–1185. https://doi.org/10.1007/978-94-007-0753-5_516

[36] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. 2016. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* 16, 5 (2016), 646.

[37] Emmanuel Mignot. 2008. Why we sleep: the temporal organization of recovery. *PLoS biology* 6, 4 (2008).

[38] Audrey Millar, Colin A Espie, and Jan Scott. 2004. The sleep of remitted bipolar outpatients: a controlled naturalistic study using actigraphy. *Journal of affective disorders* 80, 2-3 (2004), 145–153.

[39] Raeanne C Moore, Joel Swendsen, and Colin A Depp. 2017. Applications for self-administered mobile cognitive assessments in clinical research: A systematic review. *International journal of methods in psychiatric research* 26, 4 (2017), e1562.

[40] Kevin R Murphy and Charles O Davidshofer. 1988. Psychological testing. *Principles, and Applications, Englewood Cliffs* (1988).

[41] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691–1696.

[42] Reshmi Pal, John Mendelson, Odile Clavier, Mathew J Baggott, Jeremy Coyle, and Gantt P Galloway. 2016. Development and testing of a smartphone-based cognitive/neuropsychological evaluation system for substance abusers. *Journal of psychoactive drugs* 48, 4 (2016), 288–294.

[43] Indrajeet Patil. 2018. *ggstatsplot: "ggplot2" Based Plots with Statistical Details.* https://doi.org/10.5281/zenodo.2074621

[44] Ralph M Reitan. 1958. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and motor skills* 8, 3 (1958), 271–276.

[45] Paulo Marcos Brasil Rocha, Fernando Silva Neves, and Humberto Corrêa. 2013. Significant sleep disturbances in euthymic bipolar patients. *Comprehensive psychiatry* 54, 7 (2013), 1003–1008.

[46] Manuela Russo, Katie Mahon, Megan Shanahan, Elizabeth Ramjas, Carly Solon, Shaun M Purcell, and Katherine E Burdick. 2015. The relationship between sleep quality and neurocognition in bipolar disorder. *Journal of affective disorders* 187 (2015), 156–162.

[47] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[48] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.

[49] Christina Schmidt, Fabienne Collette, Christian Cajochen, and Philippe Peigneux. 2007. A time to think: circadian rhythms in human cognition. *Cognitive neuropsychology* 24, 7 (2007), 755–789.

[50] Martin J Sliwinski, Jacqueline A Mogle, Jinshil Hyun, Elizabeth Munoz, Joshua M Smyth, and Richard B Lipton. 2018. Reliability and validity of ambulatory cognitive assessments. *Assessment* 25, 1 (2018), 14–30.

[51] Ingo Steinwart and Andreas Christmann. 2008. *Support vector machines.* Springer Science & Business Media.

[52] Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine* (1994).

[53] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4, 4 (1994), 295–312.

[54] Alejandro Szmulewicz, Marina P Valerio, and Diego J Martino. 2019. Longitudinal analysis of cognitive performances in recent-onset and late-life Bipolar Disorder: A systematic review and meta-analysis. *Bipolar disorders* (2019).

[55] Zoë Tieges, Antaine Stíobhairt, Katie Scott, Klaudia Suchorab, Alexander Weir, Stuart Parks, Susan Shenkin, and Alasdair MacLullich. 2015. Development of a smartphone application for the objective detection of attentional deficits in delirium. *International psychogeriatrics* 27, 8 (2015), 1251–1262.

[56] Corrie Timmers, Anne Maeghs, Michiel Vestjens, Charlie Bonnemayer, Huub Hamers, and Arjan Blokland. 2014. Ambulant cognitive assessment using a smartphone. *Applied Neuropsychology: Adult* 21, 2 (2014), 136–142.

[57] John Torous, JP Onnela, and Matcheri Keshavan. 2017. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. *Translational psychiatry* 7, 3 (2017), e1053–e1053.

[58] John Torous, Patrick Staples, Ian Barnett, Luis R Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela. 2018. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *NPJ digital medicine* 1, 1 (2018), 1–9.

[59] Eirini Tsitsipa and Konstantinos N Fountoulakis. 2015. The neurocognitive functioning in bipolar disorder: a systematic review of data. *Annals of general psychiatry* 14, 1 (2015), 42.

[60] Céline Vetter, Myriam Juda, and Till Roenneberg. 2012. The influence of internal time, time awake, and sleep duration on cognitive performance in shiftworkers. *Chronobiology international* 29, 8 (2012), 1127–1138.

[61] Robert West, Kelly J Murphy, Maria L Armilio, Fergus IM Craik, and Donald T Stuss. 2002. Effects of time of day on age differences in working memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 57, 1 (2002), P3–P10.

[62] Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

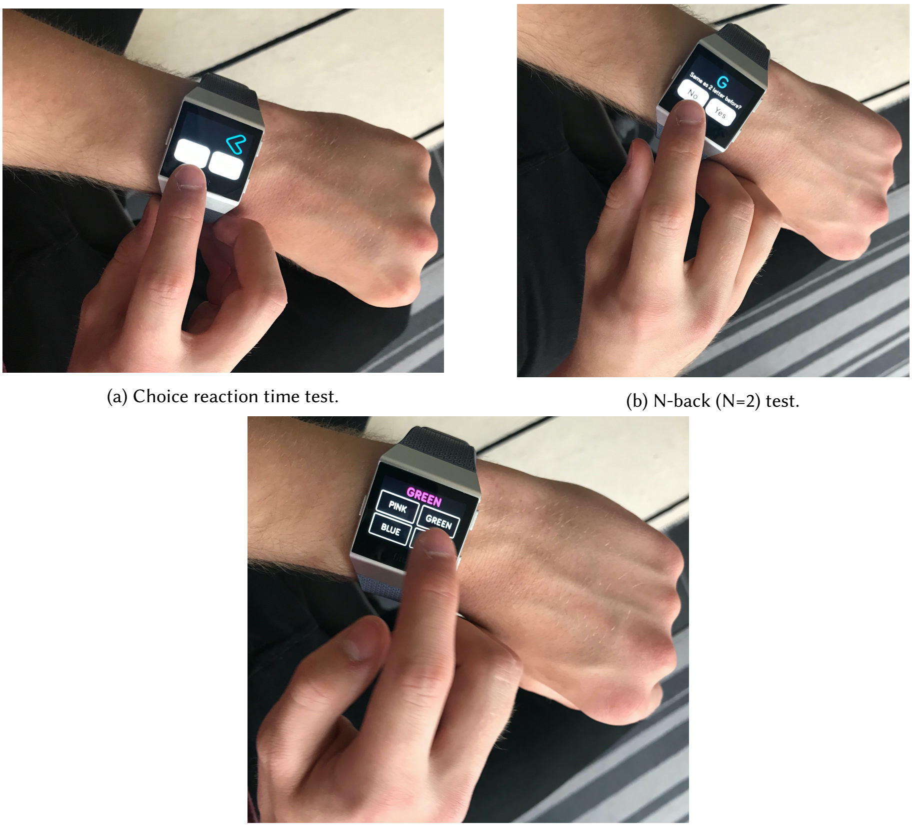
[63] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[64] Bin Xie and Gavriel Salvendy. 2000. Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments. *Work & stress* 14, 1 (2000), 74–99.

[65] Yan Zhang and Xindong Wu. 2010. Integrating induction and deduction for noisy data mining. *Information Sciences* 180, 14 (2010), 2663–2673.

A SMARTWATCH-BASED COGNITIVE TESTS

Figures 11a to 11c show snapshots of the smartwatch-based cognitive tests. The choice reaction time test has 40 arrows that appear on either right or left side of the watch screen. The arrows are right-hand or left-hand. The participants should select the direction of each arrow by tapping on either the right-hand or left-hand rectangle (touch button) (see Figure 11a). The N-back test has three difficulty levels determined by the N value. This test shows a sequence of 40 letters one by one. Figure 11b shows a 2-back task, thus, the participant should tap on 'Yes', if the letter G had appeared 2 letters back in the sequence. The Stroop test displays 30 color names one by one where each stimuli is either congruent or incongruent. Figure 11c shows an incongruent stimuli as "Green" is written in a pink color.



(a) Choice reaction time test.

(b) N-back (N=2) test.

(c) Stroop color-word test.

Fig. 11. Smartwatch-based Cognitive Assessment.

APPENDIX A

Screen for Cognitive Impairment in Psychiatry

SCIP FORM 3

SCREEN FOR COGNITIVE IMPAIRMENT IN PSYCHIATRY (SCIP)

Scot E. Purdon, Ph.D., Clinical Professor of Psychiatry, University of Alberta; spurdon@ualberta.ca

1. List learning test: Read the list of 10 words at 3 seconds per word. Test free recall. Repeat 2 more times. At the end of trial 3 let participant know they will be asked to recall the list again later.

	Desert	Face	Letter	Bed	Machine	Milk	Helmet	Sailor	Horse	Nail	Σ/10	
Tr. 1												
Tr. 2												
Tr. 3												Σ/30 =

2. Consonant repetition test (Read each set of three letters. Have the subject count backwards from the start # for the seconds under delay for each item, and then recall letters. Any order is fine):

Stimulus	Start #	Delay	Response		Stimulu s	Start #	Delay	Response	
D-L-H					Z-Q-M	49	3		
M-S-R					B-X-K	67	18		
P-H-Q	39	9			N-F-P	128	9		
X-C-D	177	18			C-T-J	40	3		Σ/24 =

3. Verbal fluency test. Allow 30 seconds to generate words beginning with each letter.

Stimulus	Response	
F		
R		Σ =

4. Delayed list learning: Ask the subject to recall the earlier words; do not repeat the list.

	Desert	Face	Letter	Bed	Machine	Milk	Helmet	Sailor	Horse	Nail		
Tr. 4											Σ /10	t4/t3 * 100

---FOLD HERE---

5. Visuomotor tracking test: After practice items, allow 30 seconds to complete left to right and top to bottom.

M	F	X	D	W	J
--	..-.	-..-	-..	.--	.---

Practice						Test			
W	D	X	J	M	F	X	M	W	
F	J	D	W	D	M	J	X	F	
M	X	J	W	D	F	X	J	F	
D	W	M	F	X	W	M	F	J	Σ/30=

SCORING SUMMARY: For each sub-test, divide the difference between observed from predicted scores and divide by the standard deviation (n=185, 1st year college sample, IQ approx. 110): Z-Scores=((Score-Mean)/SD). M±SD for VLT_I=23.59±2.87, WMT=20.66±2.45, VFT=17.44-4.74, VLT_D=7.65±1.90, PST=14.26±2.25.

Subject Name (First, Last):__________________________Gender: ____ Examiner::______________
DOB (d/m/y):__________________ Test Date (d/m/y):_______________ Time of test:____________
IQ estimate (indicate PPVT, NART, WAIS):_____________Education (years):___________Handedness:____

APPENDIX B
Usability Questionnaire

These questions are selected from Mobile Application Rating Scale (MARS) questionnaire [82].

Usability Evaluation

* Required

Aesthetics

1. **Is arrangement and size of buttons/icons/content on the screen appropriate?** *

   Mark only one oval.

   ○ Very bad design, cluttered, some options impossible to select/locate/see/read device display not optimised

   ○ Bad design, random, unclear, some options difficult to select/locate/see/read

   ○ Satisfactory, few problems with selecting/locating/seeing/reading items or with minor screen-size problems

   ○ Mostly clear, able to select/locate/see/read items

   ○ Professional, simple, clear, orderly, logically organised, device display optimised. Every design component has a purpose

2. **How high is the quality/resolution of graphics used for buttons/icons/content?** *

   Mark only one oval.

   ○ Graphics appear amateur, very poor visual design - disproportionate, completely stylistically inconsistent

   ○ Low quality/low resolution graphics; low quality visual design – disproportionate, stylistically inconsistent

   ○ Moderate quality graphics and visual design (generally consistent in style)

   ○ High quality/resolution graphics and visual design – mostly proportionate, stylistically consistent

   ○ Very high quality/resolution graphics and visual design - proportionate, stylistically consistent throughout

3. **How good does the app look?** *

   Mark only one oval.

   ○ No visual appeal, unpleasant to look at, poorly designed, clashing/mismatched colours

   ○ Little visual appeal – poorly designed, bad use of colour, visually boring

   ○ Some visual appeal – average, neither pleasant, nor unpleasant

   ○ High level of visual appeal – seamless graphics – consistent and professionally designed

   ○ As above + very attractive, memorable, stands out; use of colour enhances app features/menus

Functionality

4. **Performance: How accurately/fast do the app features (functions) and components (buttons/menus) work?** *

Mark only one oval.

○ App is broken; no/insufficient/inaccurate response (e.g. crashes/bugs/broken features, etc.)

○ Some functions work, but lagging or contains major technical problems

○ App works overall. Some technical problems need fixing/Slow at times

○ Mostly functional with minor/negligible problems

○ Perfect/timely response; no technical bugs found/contains a 'loading time left' indicator

5. **Ease of use: How easy is it to learn how to use the app; how clear are the menu labels/icons and instructions?** *

Mark only one oval.

○ No/limited instructions; menu labels/icons are confusing; complicated

○ Useable after a lot of time/effort

○ Useable after some time/effort

○ Easy to learn how to use the app (or has clear instructions)

○ Able to use app immediately; intuitive; simple

Information

6. **Quality of information: Are instructions content correct, well written, and relevant to the goal/topic of the app?** *

Mark only one oval.

○ Irrelevant/inappropriate/incoherent/incorrect

○ Poor. Barely relevant/appropriate/coherent/may be incorrect

○ Moderately relevant/appropriate/coherent/and appears correct

○ Relevant/appropriate/coherent/correct

○ Highly relevant, appropriate, coherent, and correct

7. **Quantity of information: Is the extent coverage within the scope of the app and comprehensive but concise?** *

Mark only one oval.

○ Minimal or overwhelming

○ Insufficient or possibly overwhelming

○ OK but not comprehensive or concise

○ Offers a broad range of information, has some gaps or unnecessary detail; or has no links to more information and resources

○ Comprehensive and concise; contains links to more information and resources

Bibliography

[1] Saeed Abdullah et al. “Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2016, pages 178–189.

[2] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. “User-centered design”. In: *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37.4 (2004), pages 445–456.

[3] Sanchit Aggarwal. “Modern Web-Development using ReactJS”. In: *International Journal of Recent Research Aspects* 5 (2018), pages 133–137.

[4] Michèle Allard et al. “Mobile technologies in the early detection of cognitive decline”. In: *PLoS One* 9.12 (2014), e112197.

[5] Naomi S Altman. “An introduction to kernel and nearest-neighbor nonparametric regression”. In: *The American Statistician* 46.3 (1992), pages 175–185.

[6] Alexandra S Atkins et al. “Validation of the tablet-administered Brief Assessment of Cognition (BAC App)”. In: *Schizophrenia research* 181 (2017), pages 100–106.

[7] Patrick Biernacki and Dan Waldorf. “Snowball sampling: Problems and techniques of chain referral sampling”. In: *Sociological methods & research* 10.2 (1981), pages 141–163.

[8] E Bora and A Özerdem. “Meta-analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia”. In: *Psychological medicine* 47.16 (2017), pages 2753–2766.

[9] Robert M Brouillette et al. “Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly”. In: *PloS one* 8.6 (2013), e65925.

[10] LTD Cambridge Cognition. *Cambridge Neuropsychological Test Automated Battery (CANTAB)*. 1996.

[11] Naomi Chaytor and Maureen Schmitter-Edgecombe. “The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills”. In: *Neuropsychology review* 13.4 (2003), pages 181–197.

[12] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pages 785–794.

[13] Stephanie Hui-Wen Chuah et al. “Wearable technologies: The role of usefulness and visibility in smartwatch adoption”. In: *Computers in Human Behavior* 65 (2016), pages 276–284.

[14] Cambridge Cognition. *6 cognitive tests for research of depression and mood disorders*. 2015.

[15] Francesca Cormack et al. “Wearable Technology for High-Frequency Cognitive and Mood Assessment in Major Depressive Disorder: Longitudinal Observational Study”. In: *JMIR Mental Health* 6.11 (2019), e12814.

[16] Paul Dagum. “Digital biomarkers of cognitive function”. In: *NPJ digital medicine* 1.1 (2018), pages 1–3.

[17] NEM Daniëls et al. “Digital assessment of working memory and processing speed in everyday life: Feasibility, validation, and lessons-learned”. In: *Internet Interventions* 19 (2020), page 100300.

[18] Margaret T Davis et al. “Preliminary evidence concerning the pattern and magnitude of cognitive dysfunction in major depressive disorder using cogstate measures”. In: *Journal of affective disorders* 218 (2017), pages 82–85.

[19] Faith Dickerson et al. “Cognitive functioning in schizophrenia and bipolar disorder: comparison of performance on the Repeatable Battery for the Assessment of Neuropsychological Status”. In: *Psychiatry research* 129.1 (2004), pages 45–53.

[20] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. “Building cognition-aware systems: A mobile toolkit for extracting time-of-day fluctuations of cognitive performance”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), page 47.

[21] Anna C Domen et al. “The validation of a new online cognitive assessment tool: The MyCognition Quotient”. In: *International journal of methods in psychiatric research* 28.3 (2019), e1775.

[22] Franciscus Cornelis Donders. “On the speed of mental processes”. In: *Acta psychologica* 30 (1969), pages 412–431.

[23] Joseph S Dumas, Joseph S Dumas, and Janice Redish. *A practical guide to usability testing*. Intellect books, 1999.

[24] Tzvi Dwolatzky. “The Mindstreams computerized assessment battery for cognitive impairment and dementia”. In: *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*. 2011, pages 1–4.

[25] Maria Faurholt-Jepsen et al. “Objective smartphone data as a potential diagnostic marker of bipolar disorder”. In: *Australian & New Zealand Journal of Psychiatry* 53.2 (2019), pages 119–128.

[26] Maria Faurholt-Jepsen et al. “Smartphone data as an electronic biomarker of illness activity in bipolar disorder”. In: *Bipolar disorders* 17.7 (2015), pages 715–728.

[27] Simon Folkard and Timothy H Monk. “The measurement of circadian rhythms in psychological functions”. In: *Chronobiotechnology and chronobiological engineering*. Springer, 1987, pages 189–201.

[28] Simon Folkard and Timothy H Monk. “Time of day and processing strategy in free recall”. In: *Quarterly Journal of Experimental Psychology* 31.3 (1979), pages 461–475.

[29] Simon Folkard et al. “Independence of the circadian rhythm in alertness from the sleep/wake cycle”. In: *Nature* 313.6004 (1985), pages 678–679.

[30] Lars Frings et al. “Early detection of behavioral side effects of antiepileptic treatment using handheld computers”. In: *Epilepsy & Behavior* 13.2 (2008), pages 402–406.

[31] C Thomas Gualtieri and Lynda G Johnson. “Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs”. In: *Archives of Clinical Neuropsychology* 21.7 (2006), pages 623–643.

[32] Pegah Hafiz and Jakob E Bardram. “Design and formative evaluation of cognitive assessment apps for wearable technologies”. In: *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, pages 1162–1165.

[33] Pegah Hafiz and Jakob E Bardram. “The Ubiquitous Cognitive Assessment Tool for Smartwatches: Design, Implementation, and Evaluation Study”. In: *JMIR mHealth and uHealth* 8.6 (2020), e17506.

[34] Pegah Hafiz, Alban Maxhuni, and Jakob E Bardram. “Analysis of Perceived Human Factors and Participants’ Demographics during a Cognitive Assessment Study with a Smartwatch”. In: *2020 IEEE International Conference on Healthcare Informatics (ICHI)*. 2020.

[35] Pegah Hafiz et al. “Design and implementation of a web-based application to assess cognitive impairment in affective disorder”. In: *Proceedings of the 2018 International Conference on Digital Health*. 2018, pages 154–155.

[36] Pegah Hafiz et al. “The Internet-Based Cognitive Assessment Tool: System Design and Feasibility Study”. In: *JMIR formative research* 3.3 (2019), e13898.

[37] Pegah Hafiz et al. “Wearable Computing Technology for Assessment of Cognitive Functioning of Bipolar Patients and Healthy Controls”. 2020.

[38] Åsa Hammar and Guro Årdal. "Cognitive functioning in major depression-a summary". In: *Frontiers in human neuroscience* 3 (2009), page 26.

[39] John E Harrison et al. "Stability, reliability, and validity of the THINC-it screening tool for cognitive impairment in depression: A psychometric exploration in healthy volunteers". In: *International journal of methods in psychiatric research* 27.3 (2018), e1736.

[40] Sandra G Hart and Lowell E Staveland. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In: *Advances in psychology*. Volume 52. Elsevier, 1988, pages 139–183.

[41] Yuichi Higashiyama et al. "The neural basis of typewriting: A functional MRI study". In: *PloS one* 10.7 (2015).

[42] Shan Hung et al. "Smartphone-based ecological momentary assessment for Chinese patients with depression: An exploratory study in Taiwan". In: *Asian journal of psychiatry* 23 (2016), pages 131–136.

[43] iHope. *iHope Network*. Last accessed May 2020. URL: `https://www.ihopenetwork.com`.

[44] Monique WM Jaspers et al. "The think aloud method: a guide to user interface design". In: *International journal of medical informatics* 73.11-12 (2004), pages 781–795.

[45] Andrew Jones et al. "Do daily fluctuations in inhibitory control predict alcohol consumption? An ecological momentary assessment study". In: *Psychopharmacology* 235.5 (2018), pages 1487–1496.

[46] Susan Jongstra et al. "Cognitive testing in people at increased risk of dementia using a smartphone app: the iVitality proof-of-principle study". In: *JMIR mHealth and uHealth* 5.5 (2017), e68.

[47] Slava Kalyuga. "Cognitive load theory: How many types of load does it really need?" In: *Educational Psychology Review* 23.1 (2011), pages 1–19.

[48] Semion Kertzman et al. "Antipsychotic treatment in schizophrenia: the role of computerized neuropsychological assessment". In: *Israel journal of psychiatry and related sciences* 45.2 (2008), page 114.

[49] Wayne K Kirchner. "Age differences in short-term retention of rapidly changing information." In: *Journal of experimental psychology* 55.4 (1958), page 352.

[50] Alexandra König et al. "Fully automatic speech-based analysis of the semantic verbal fluency task". In: *Dementia and geriatric cognitive disorders* 45.3-4 (2018), pages 198–209.

[51] Tania Lara, Juan Antonio Madrid, and Ángel Correa. "The vigilance decrement in executive function is attenuated when individual chronotypes perform at their optimal time of day". In: *PloS one* 9.2 (2014), e88820.

[52] James R Lewis. “IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use”. In: *International Journal of Human-Computer Interaction* 7.1 (1995), pages 57–78.

[53] Muriel Deutsch Lezak et al. *Neuropsychological assessment.* Oxford University Press, USA, 1995.

[54] Andy Liaw, Matthew Wiener, et al. “Classification and regression by randomForest”. In: *R news* 2.3 (2002), pages 18–22.

[55] Wei-Ling Lin and Grace Yao. “Concurrent Validity”. In: *Encyclopedia of Quality of Life and Well-Being Research.* Edited by Alex C. Michalos. Dordrecht: Springer Netherlands, 2014, pages 1184–1185. ISBN: 978-94-007-0753-5. DOI: `10.1007/978-94-007-0753-5_516`. URL: `https://doi.org/10.1007/978-94-007-0753-5_516`.

[56] MR Lovell et al. “ImPACT: Immediate post-concussion assessment and cognitive testing”. In: *Pittsburgh, PA: NeuroHealth Systems, LLC* (2000).

[57] G Lyon and Norman A Krasnegor. *Attention, memory, and executive function.* Paul H Brookes Publishing Co., 1996.

[58] Wendy E Mackay and Anne-Laure Fayard. “HCI, natural science and design: a framework for triangulation across disciplines”. In: *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques.* 1997, pages 223–234.

[59] AJM Mara van Osch et al. “User preferences and usability of iVitality: optimizing an innovative online research platform for home-based health monitoring”. In: *Patient preference and adherence* 9 (2015), page 857.

[60] Thomas D Marcotte and Igor Grant. *Neuropsychology of everyday functioning.* Guilford Press, 2009.

[61] Roger S McIntyre et al. “The THINC-Integrated Tool (THINC-it) Screening Assessment for Cognitive Dysfunction: Validation in Patients With Major Depressive Disorder.” In: *The Journal of clinical psychiatry* 78.7 (2017), pages 873–881.

[62] Kathleen R Merikangas et al. “Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative”. In: *Archives of general psychiatry* 68.3 (2011), pages 241–251.

[63] Timothy H Monk et al. “Task variables determine which biological clock controls circadian rhythms in human performance”. In: *Nature* 304.5926 (1983), page 543.

[64] Kevin R Murphy and Charles O Davidshofer. “Psychological testing”. In: *Principles, and Applications, Englewood Cliffs* (1988).

[65] Shahriar Nirjon et al. “MOBI-COG: a mobile application for instant screening of dementia using the mini-cog test”. In: *Proceedings of the Wireless Health 2014 on National Institutes of Health.* 2014, pages 1–7.

[66] Jukka-Pekka Onnela and Scott L Rauch. “Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health”. In: *Neuropsychopharmacology* 41.7 (2016), pages 1691–1696.

[67] Serguei VS Pakhomov et al. “Using automatic speech recognition to assess spoken responses to cognitive tests of semantic verbal fluency”. In: *Speech communication* 75 (2015), pages 14–26.

[68] Reshmi Pal et al. “Development and testing of a smartphone-based cognitive/neuropsychological evaluation system for substance abusers”. In: *Journal of psychoactive drugs* 48.4 (2016), pages 288–294.

[69] Scot E Purdon and R Psych. “THE SCREEN FOR COGNITIVE IMPAIRMENT IN PSYCHIATRY”. In: (2005).

[70] Trevor W Robbins et al. “Cambridge Neuropsychological Test Automated Battery (CANTAB): a factor analytic study of a large sample of normal elderly volunteers”. In: *Dementia and geriatric cognitive disorders* 5.5 (1994), pages 266–281.

[71] Yvonne Rogers et al. “Why it’s worth the hassle: The value of in-situ studies when designing ubicomp”. In: *International Conference on Ubiquitous Computing*. Springer. 2007, pages 336–353.

[72] Benjamin J Sadock and Virginia A Sadock. *Kaplan and Sadock’s synopsis of psychiatry: Behavioral sciences/clinical psychiatry*. Lippincott Williams & Wilkins, 2002.

[73] Sohrab Saeb et al. “Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study”. In: *Journal of medical Internet research* 17.7 (2015), e175.

[74] Sohrab Saeb et al. “The relationship between mobile phone location sensor data and depressive symptom severity”. In: *PeerJ* 4 (2016), e2537.

[75] Peter Sandrini. “Website localization and translation”. In: *EU-High-Level Scientific Conference Series MuTra*. 2005, pages 131–138.

[76] Judith Saxton et al. “Computer assessment of mild cognitive impairment”. In: *Postgraduate medicine* 121.2 (2009), pages 177–185.

[77] Christina Schmidt et al. “A time to think: circadian rhythms in human cognition”. In: *Cognitive neuropsychology* 24.7 (2007), pages 755–789.

[78] Martin J Sliwinski et al. “Reliability and validity of ambulatory cognitive assessments”. In: *Assessment* 25.1 (2018), pages 14–30.

[79] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[80] Gijsbert Stoet. “PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments”. In: *Teaching of Psychology* 44.1 (2017), pages 24–31.

[81] Gijsbert Stoet. “PsyToolkit: A software package for programming psychological experiments using Linux”. In: *Behavior Research Methods* 42.4 (2010), pages 1096–1104.

[82] Stoyan R Stoyanov et al. “Mobile app rating scale: a new tool for assessing the quality of health mobile apps”. In: *JMIR mHealth and uHealth* 3.1 (2015), e27.

[83] Esther Strauss, Elisabeth MS Sherman, Otfried Spreen, et al. *A compendium of neuropsychological tests: Administration, norms, and commentary.* American Chemical Society, 2006.

[84] J Ridley Stroop. “Studies of interference in serial verbal reactions.” In: *Journal of experimental psychology* 18.6 (1935), page 643.

[85] John A Sweeney, Julie A Kmiec, and David J Kupfer. “Neuropsychologic impairments in bipolar and unipolar mood disorders on the CANTAB neurocognitive battery”. In: *Biological psychiatry* 48.7 (2000), pages 674–684.

[86] John Sweller. “Cognitive load theory, learning difficulty, and instructional design”. In: *Learning and instruction* 4.4 (1994), pages 295–312.

[87] Alejandro Szmulewicz, Marina P Valerio, and Diego J Martino. “Longitudinal analysis of cognitive performances in recent-onset and late-life Bipolar Disorder: A systematic review and meta-analysis”. In: *Bipolar disorders* (2019).

[88] Zoë Tieges et al. “Development of a smartphone application for the objective detection of attentional deficits in delirium”. In: *International psychogeriatrics* 27.8 (2015), pages 1251–1262.

[89] Corrie Timmers et al. “Ambulant cognitive assessment using a smartphone”. In: *Applied Neuropsychology: Adult* 21.2 (2014), pages 136–142.

[90] Brian Tiplady et al. “Alcohol and cognitive function: assessment in everyday life and laboratory settings using mobile phones”. In: *Alcoholism: Clinical and Experimental Research* 33.12 (2009), pages 2094–2102.

[91] Jane B Tornatore et al. “Self-administered screening for mild cognitive impairment: initial validation of a computerized test battery”. In: *The Journal of neuropsychiatry and clinical neurosciences* 17.1 (2005), pages 98–105.

[92] László Tóth et al. “A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech”. In: *Current Alzheimer Research* 15.2 (2018), pages 130–138.

[93] Johannes Tröger et al. “Telephone-Based Dementia Screening I: Automated Semantic Verbal Fluency Assessment”. In: *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare.* 2018, pages 59–66.

[94] Timothy J Trull and Ulrich Ebner-Priemer. “Ambulatory assessment”. In: *Annual review of clinical psychology* 9 (2013), pages 151–176.

[95] Eirini Tsitsipa and Konstantinos N Fountoulakis. “The neurocognitive functioning in bipolar disorder: a systematic review of data”. In: *Annals of general psychiatry* 14.1 (2015), page 42.

[96] Joe H Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *Journal of the American statistical association* 58.301 (1963), pages 236–244.

[97] Petra Weiland-Fiedler et al. “Evidence for continuing neuropsychological impairments in depression”. In: *Journal of affective disorders* 82.2 (2004), pages 253–258.

[98] Sandra Weintraub et al. “Cognition assessment using the NIH Toolbox”. In: *Neurology* 80.11 Supplement 3 (2013), S54–S64.

[99] Alexander J Weir et al. “Development of Android apps for cognitive assessment of dementia and delirium”. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2014, pages 2169–2172.

[100] Robert West et al. “Effects of time of day on age differences in working memory”. In: *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 57.1 (2002), P3–P10.

[101] Roderick Westerman et al. “Computer-assisted cognitive function assessment of pilots”. In: *ADF Health* 2 (2001), pages 29–36.

[102] Robert T Wilkinson and David Houghton. “Field test of arousal: a portable reaction timer with data storage”. In: *Human factors* 24.4 (1982), pages 487–493.

[103] Bin Xie and Gavriel Salvendy. “Review and reappraisal of modelling and predicting mental workload in single-and multi-task environments”. In: *Work & stress* 14.1 (2000), pages 74–99.