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

Hafiz, Pegah; Maxhuni, Alban; Bardram, Jakob Eyvind

Published in:
Proceedings of the 8th IEEE International Conference on Healthcare Informatics

Link to article, DOI:
10.1109/ICHI48887.2020.9374342

Publication date:
2021

Document Version
Peer reviewed version

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:

- 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

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)

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.

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

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

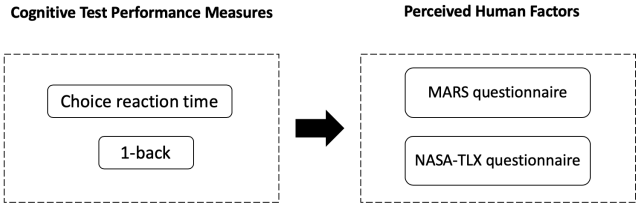

Figure (2) Schematic overview of the correlation analysis performed in this paper

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 $\approx$ 0 for identical subjects and $\approx$ 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 $\pm$ SD	**26.90 $\pm$ 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$\pm$107	39.57$\pm$0.60	36.5$\pm$5.34
1-Back	903$\pm$266	37.09$\pm$4.82	34$\pm$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 Measure	Usability Metrics		
	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 Measure	Usability Metrics		
	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-

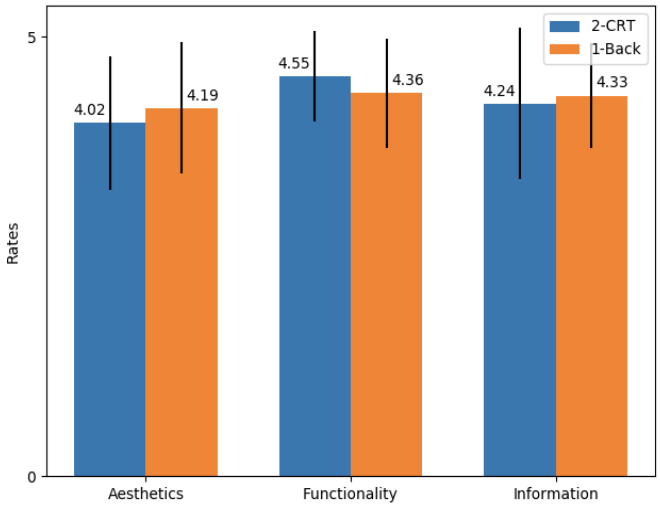

Figure (3) Usability ratings by our study participants presented separately for each cognitive test

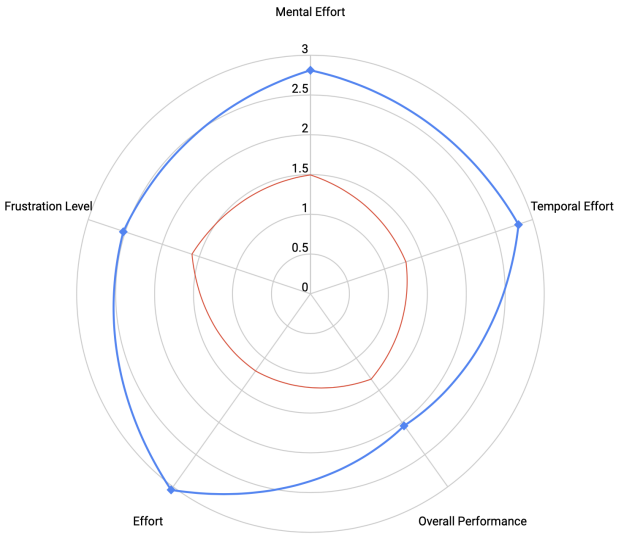

Figure (4) NASA-TLX - Sub-scales were rated by 5-point Likert scale

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

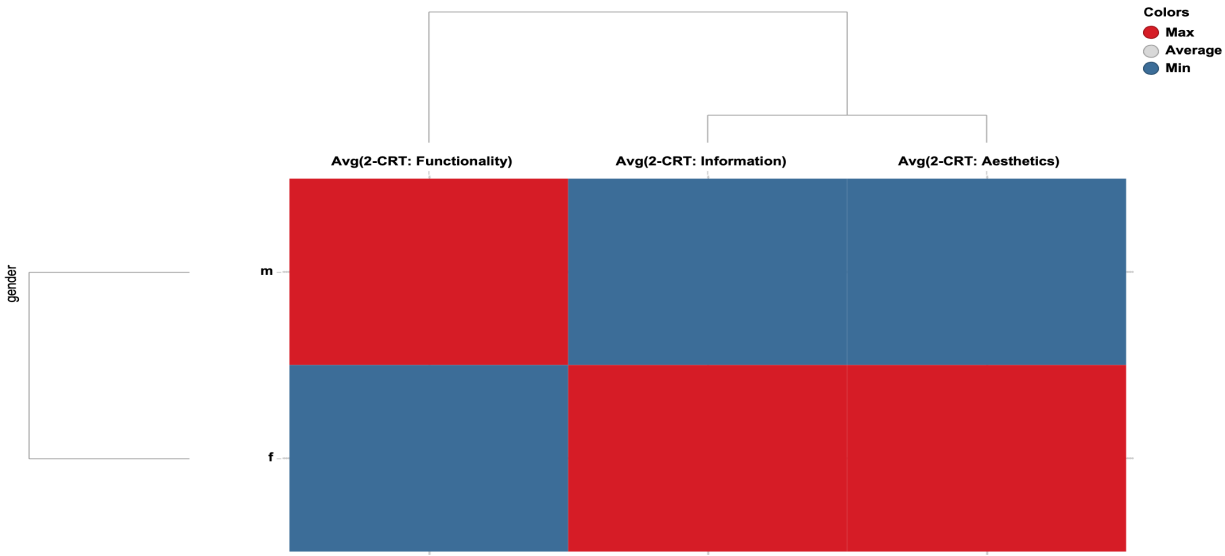

Figure (5) Clusters of participants' gender (m=male, f=female) based on their perceived usability of the two-choice reaction time test.

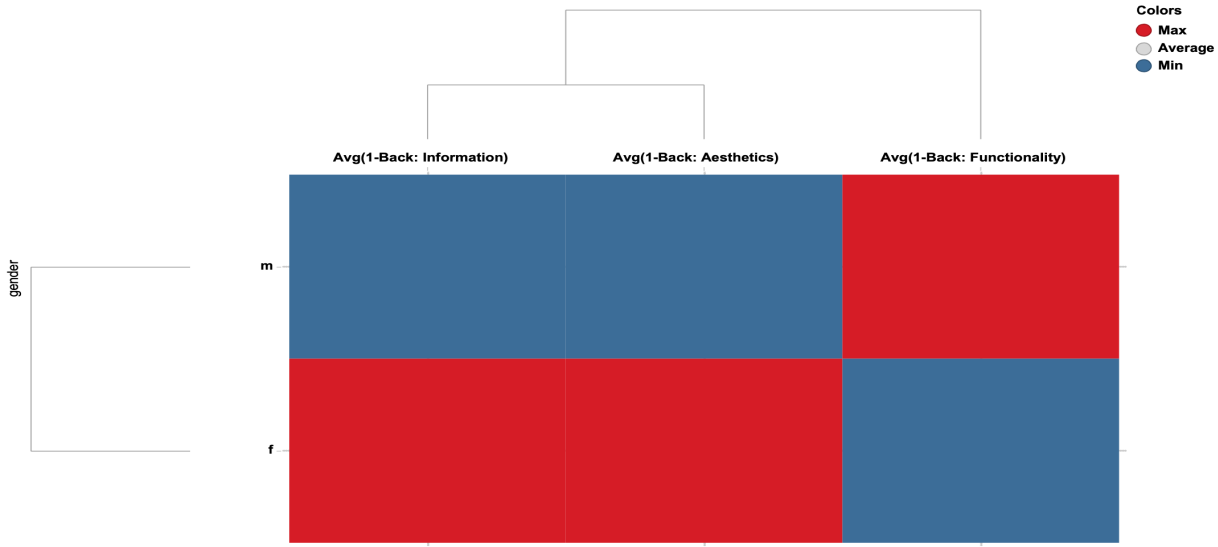

Figure (6) Clusters of participants' gender (m=male, f=female) based on their perceived usability of the 1-back App.

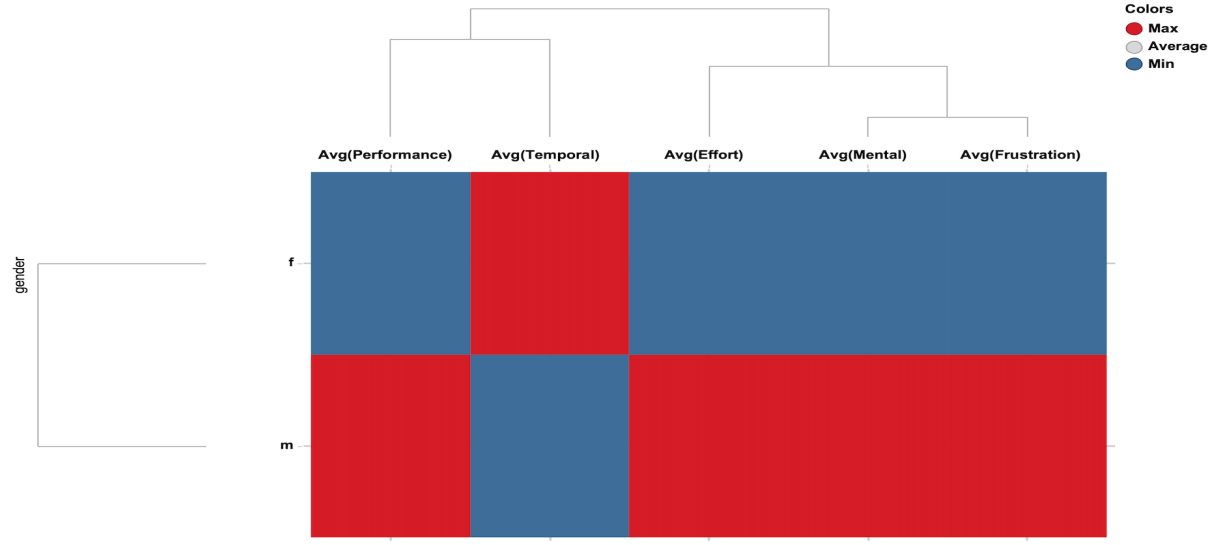

Figure (7) Clusters of participants' gender (m=male, f=female) based on their perceived cognitive load.

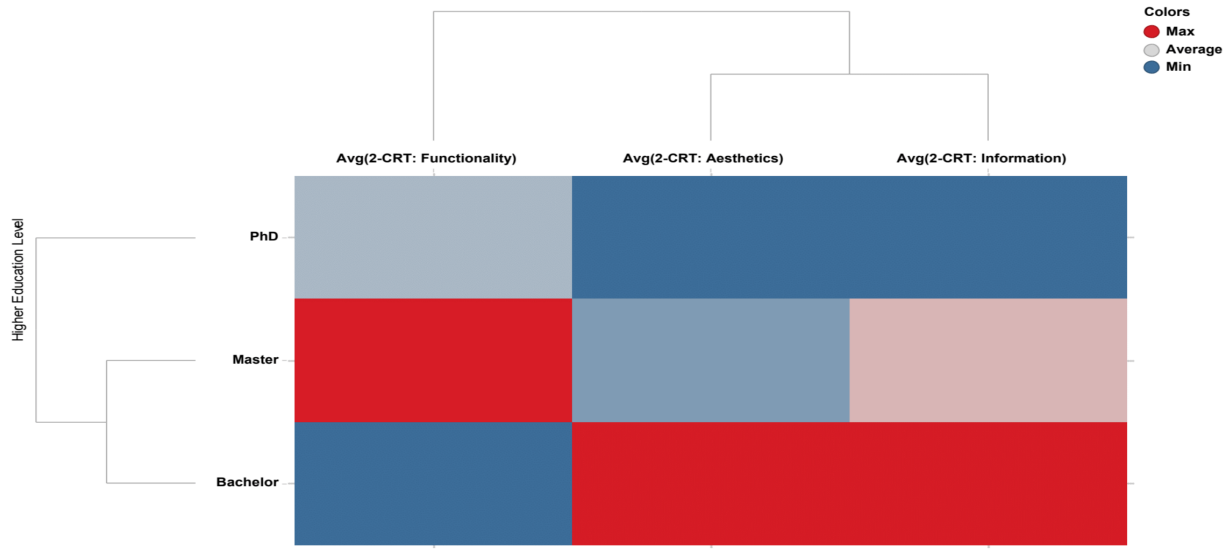

Figure (8) Clusters of participants' education based on their perceived usability of the two-choice reaction time application.

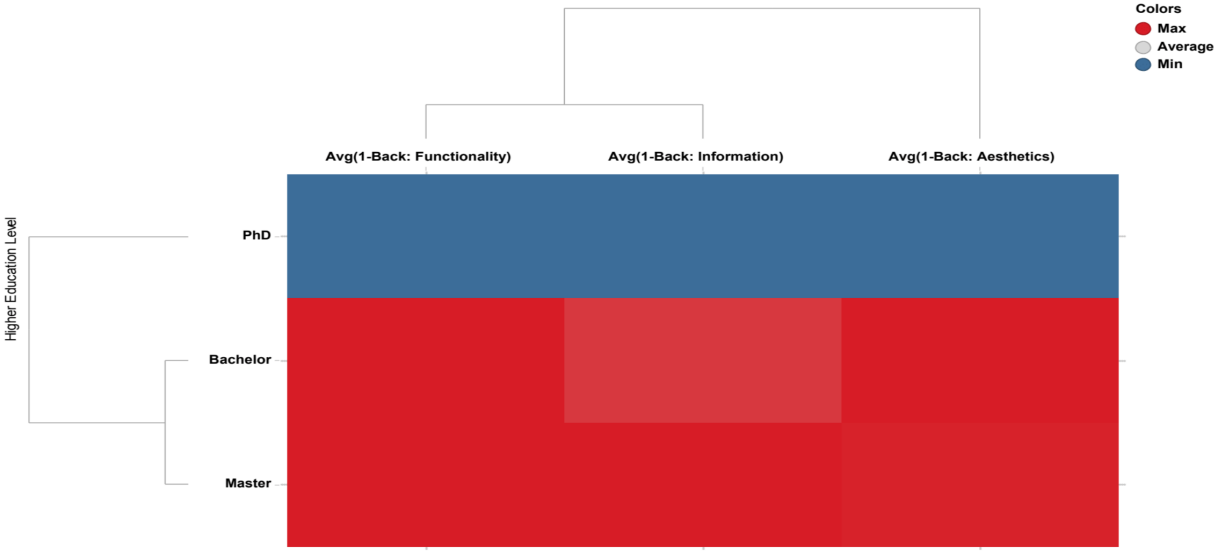

Figure (9) Clusters of participants' education based on their perceived usability of the 1-back application.

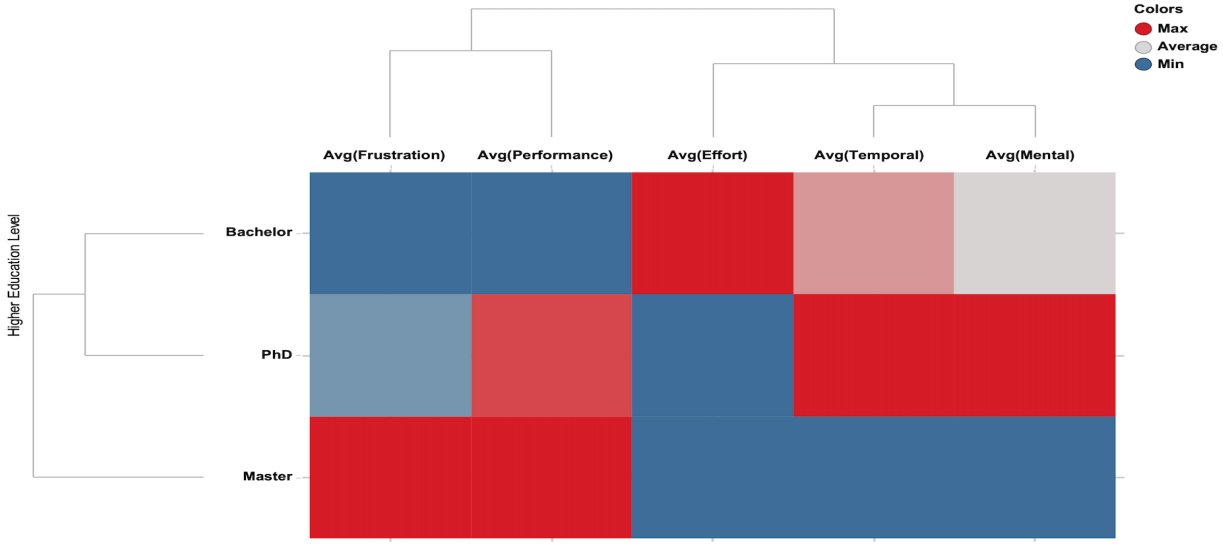

Figure (10) Clusters of participants' higher education level based on their perceived cognitive load.

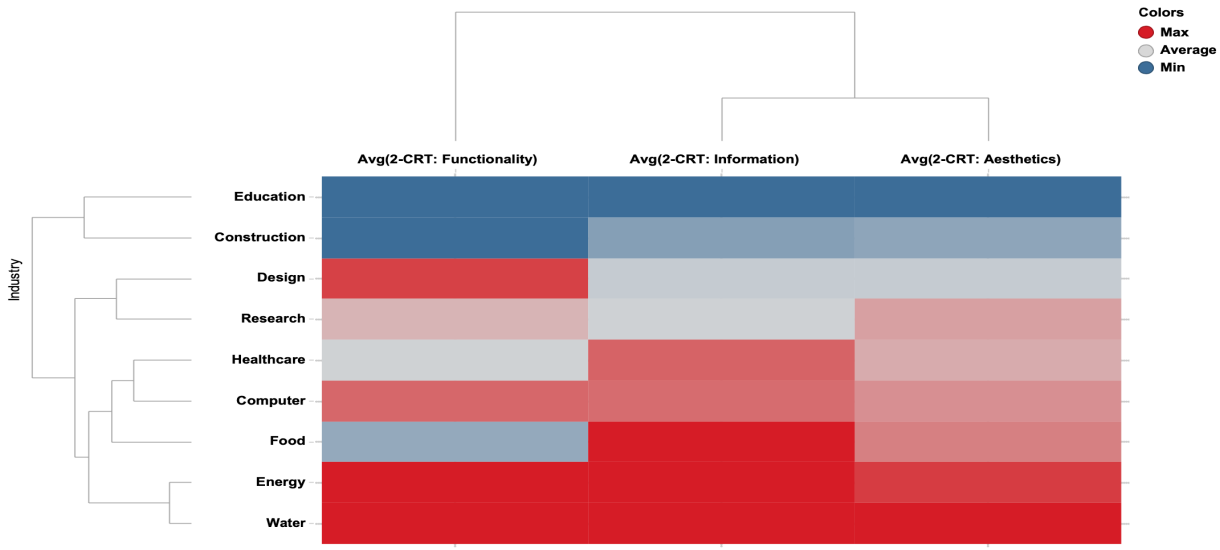

Figure (11) Clusters of participants' education or work industry based on their perceived usability of the two-choice reaction time application.

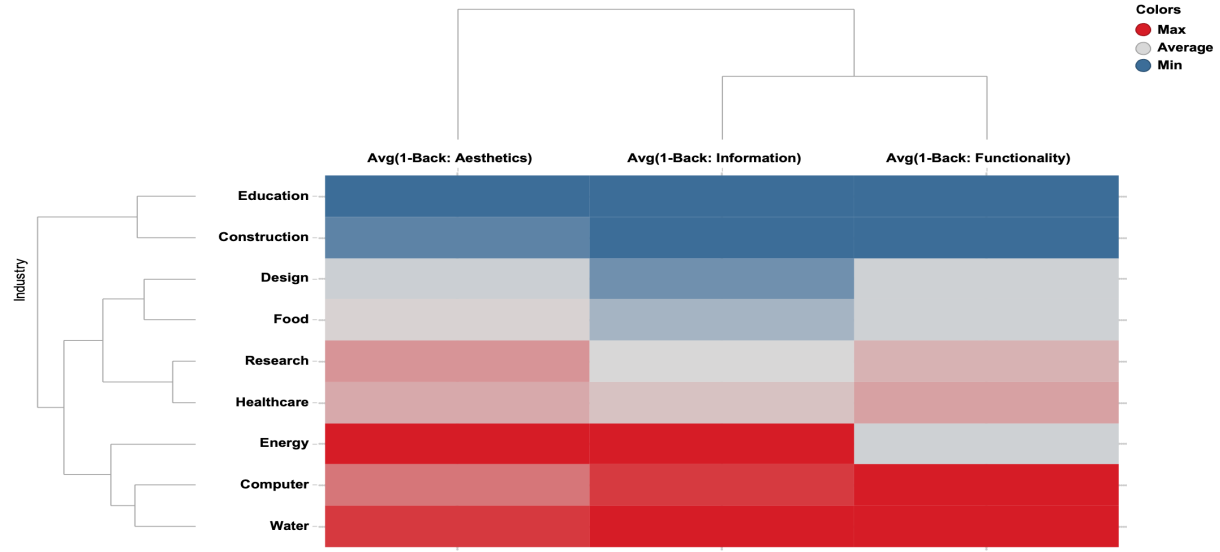

Figure (12) Clusters of participants' education or work industry based on their perceived usability of the 1-back application.

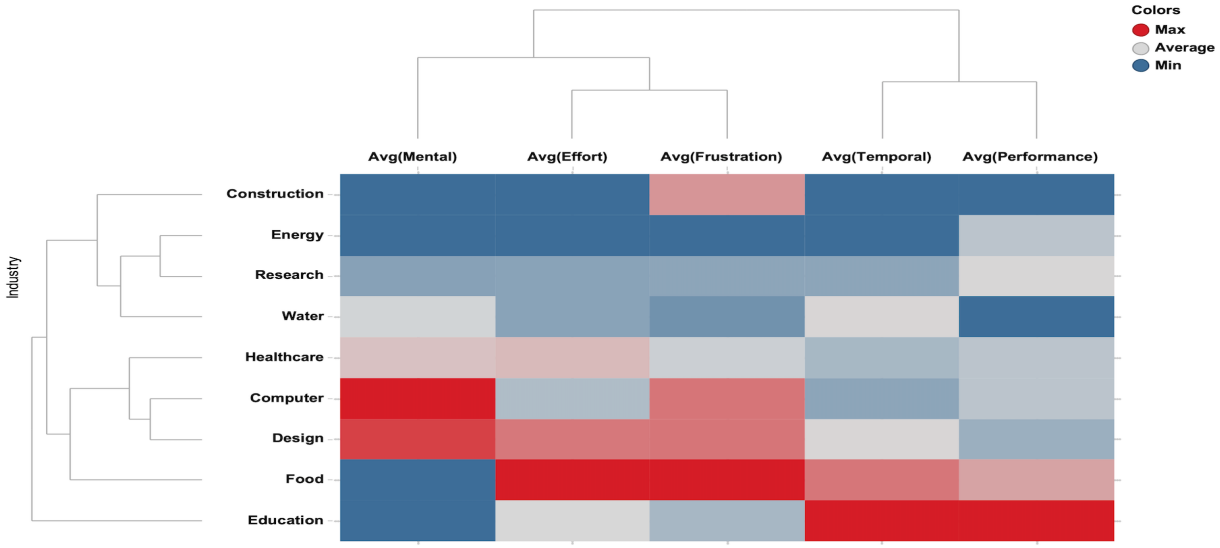

Figure (13) Clusters of participants' education or work industry based on their perceived cognitive load.

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.