9PM: A Novel Interactive 9-Peg Board for Cognitive and Physical Assessment

Maher Abujelala
Texas A&M University
College Station, Texas, USA
maher.abujelala@tamu.edu

Varun Kanal
The University of Texas at Arlington
Arlington, Texas, USA
varun.kanal@mavs.uta.edu

Akilesh Rajavenkatanarayanan
The University of Texas at Arlington
Arlington, Texas, USA
akilesh.rajavenkatanarayanan@mavs.uta.edu

Fillia Makedon
The University of Texas at Arlington
Arlington, Texas, USA
makedon@uta.edu

ABSTRACT

Cognitive assessments are a crucial part of rehabilitation in persons with a neurological disorder and vocational rehabilitation, where people need to be trained to improve their cognitive abilities. While human action involves using several cognitive skills and physical skills, most assessment systems focus on detecting or assessing either the cognitive ability or just physical ability. There is a need for a system that bridges the gap between real-world activity, which involves physical activity and cognition, and clinical tests that are tailored for a specific use. To address this need, we propose a novel interactive 9-Hole Pegboard called the 9-Peg Move (*9PM*) capable of performing both cognitive and physical assessments in the same system. The system incorporates wearable sensors to collect data for objective evaluation. Preliminary machine learning results indicate that the data collected using our system can reliably recognize cognitive factors like perceived mental effort, perceived task difficulty, and perceived interest in a task. These results are the first step toward building an automated immersive assessment system.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Health informatics**.

KEYWORDS

Physical assessment, cognitive assessment, Nine Hole Peg Test, physiological sensing, wearable sensors

ACM Reference Format:
Maher Abujelala, Varun Kanal, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2021. 9PM: A Novel Interactive 9-Peg Board for Cognitive and Physical Assessment. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3453892.3453996

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3453996

1 INTRODUCTION

Advancements in wearable sensor technology in recent years have paved the way to several body-wearable sensors like the Electrocardiogram (ECG), Electroencephalogram (EEG), Inertial Measurement Units (IMU), and electrodermal activity (EDA) for use in clinical research activities [14, 27]. We have also seen the incorporation of some of these sensors into low-cost, power-efficient devices such as activity trackers and smartwatches, making them available for a larger population of developers, researchers, and users. Data from such devices can help track movement, sleep pattern and quality, heart-rate data, and many others. Leveraging this vast amount of data can help us understand more about our bodies and capabilities. We can use it to advance healthcare and educational applications by personalizing them based on our needs.

In recent times, researchers are using sensors to help them build sensor-based smart assessment systems in healthcare [13, 18, 25]. Recent research also suggests the development of a personalized vocational, cognitive skill assessment and training system to assess human worker's cognitive skills required for assembly and manufacturing tasks [3, 35]. However, assessing the skills of an individual is not straightforward. Human experts have the ability to evaluate a person's ability intuitively using specific tests and years of experience; however, developing smart systems and algorithms can be a daunting task.

While there are specific tests to assess specific cognitive or physical abilities, real-world tasks involve working on a physical task while using their cognition. For example, an assembly line worker may need to assemble parts of a product using proper hand dexterity while requiring to remember the assembly's steps using working memory. In this paper, we discuss our effort to build a novel assessment system that makes it possible to combine physical and cognitive assessment. Our system, the 9-Peg Move (*9PM*), shown in Figure 1, can be used to build assessments that require a physical action of moving pegs from a source location (red/blue areas) to a destination (white area) based on an underlying cognitive test. We hypothesize that using such a system, researchers can build tasks that simulate the real-world and also incorporate sensors that monitor the participants for an objective measurement of the required cognitive and physical skills. To illustrate our hypothesis, we use commonly used sensors that can be used to monitor a person while working, for changes in performance due to stress, lack of sleep,

or increase in cognitive load [15]. The main contributions of the presented work are:

- The novel design of the *9PM* Board incorporating digitized versions of physical and cognitive assessments.
- A public dataset[1] for behavioral, performance, and physiological data of participants performing modified versions of standard physical and cognitive assessments.

The paper is organized as follows. In section 2, we present related work, section 3 describes the *9PM* system for Cognitive Assessment using a physical task, and in section 4 we present the experimental results and evaluation of the study. In the final section, we discuss our future research direction and conclude.

2 RELATED WORKS

After accidents or due to disorders, a person may lose their upper limb function or dexterity. To rehabilitate the loss of function and mobility, physical tests to assess dexterity are used. The "gold standard" to evaluate manual dexterity in many disorders is the Nine Hole Peg Test (NHPT) [11, 12, 24]. A traditional NHPT consists of a board with nine holes and a receptacle. This receptacle consists of nine pegs designed to fit into the holes. When instructed, the person would move the pegs one at a time from the receptacle to the hole and back multiple times. An expert monitors the user's performance who can then assess their dexterity. Other tests like the Box and Blocks test also assess the manual dexterity of the upper arm [20]. In this test, a box that is divided in the middle by a partition is used. One part of the box contains a few blocks of equal sizes. When instructed, the person has to pick the blocks from one side of the box and place it on the other side. The user performs this action multiple times while moving above the partition. An expert monitors the user's performance to assess their dexterity.

A battery of tests is available to assess human cognition. These tests assess one or more aspects of cognition. One of the most common cognitive tests is the Stroop test, which evaluates inhibition, attention, and executive function [29]. Here, the person needs to identify an incongruent stimulus in a series of stimuli. Another aspect of cognition is episodic memory. The Picture Sequence Memory Test (PSMT) evaluates this aspect of cognition [5]. This test will show the person a series of pictures that they must arrange in a specific order. List Sorting Test assesses working memory as an aspect of cognition [36]. In this test, the person has to sort a sequence based on the stimuli given. Pattern Comparison Processing Speed Test assesses processing speed [8]. This test will show the person two pictures, and they must recognize if the pictures are the same.

The above tests are used for assessing the cognitive state of both a neurotypical person or a person suffering from neurological disorders. Administering these tests under neurological disorders gives an indication of the current cognitive state of the person and indicates the progression of the disorder. While several attempts have been made to combine cognitive assessments in a physical task [19, 26], these works focus on assessing a specific cognitive ability. Research also exists to assess multiple cognitive abilities using a single system for screening [37], but this system does not incorporate a physical task. Our *9PM* system allows the researcher to create several combinations of cognitive assessments with the NHPT (physical task) that assesses a variety of cognitive skills, as explained in section 3. To the best of our knowledge, no such system exists.

[1] $https://github.com/abujelala/9PM$

Figure 1: The *9PM* Board, a Novel Modified Version of the NHPT.

3 *9PM*: AN INTERACTIVE TOOL FOR COGNITIVE ASSESSMENT USING PHYSICAL TASK

As discussed before, traditional NHPT consists of a board with nine holes and a receptacle with nine pegs. Our *9PM* system uses a novel test board, called *9PM* board, that is designed to allow the participant to perform a modified version of the NHPT with other cognitive assessments. This setup consists of three areas, red, blue, and white, each containing nine holes. While the red and the blue are dubbed as the source areas, the white area is also known as the destination area and does not have any pegs. According to the instruction provided, the cognitive tests require the user to move the pegs from the source to the destination area.

9PM incorporates the following cognitive tests; Stroop Test [32], the Wisconsin Card Sorting Test (WCST) [17] and the NIH Toolbox Picture Sequence Memory Test (PSMT) [10]. The Stroop Test assesses inhibition, attention, and executive function, and PSMT evaluates episodic memory. On the other hand, the WCST evaluates abstract reasoning and task-shifting ability [4]. It requires participants to match cards based on count, color, or shape. The matching rule changes randomly after a few rounds and the users need to figure out the rule based on trial and error. The cognitive tasks we chose not only cover a broad variety of cognitive assessments but also cover some of the most frequently used assessments in research and clinical practice [33].

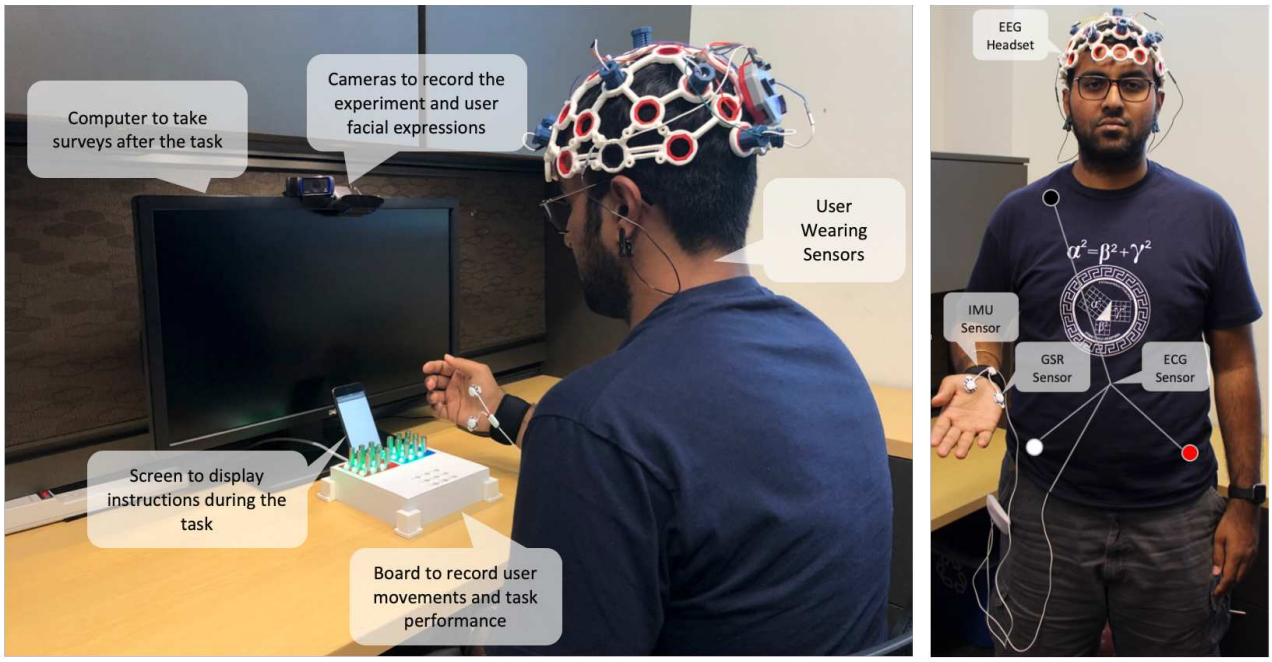

Figure 2: Experiment Setup (Left) and Sensors Placement (Right).

3.1 *9PM* Tasks

Using the *9PM* setup, the participants are asked to perform five tasks. This is accomplished by moving the pegs from one area of the board to another, thus exerting both physically and cognitively. They will move the pegs one at a time using their dominant hand. The system recognizes when a peg is picked up and when it is placed in a hole. If the participant makes a wrong move, a buzzer will sound, providing audio feedback.

In Task 1 (T1), the participant is asked to move all nine pegs to and from the white area four times in a single round, totalling 72 moves. They are asked to start from the area closest to their dominant arm. This task provided information about the participant's state when little to no cognitive effort was needed.

In Task 2 (T2), the participants are provided instructions on a screen (i.e., a smartphone). They are asked to pick a peg from either the red or the blue area and move it to the white area. The instructions are provided for each of the nine pegs.

Task 3 (T3) incorporates the Stroop test into the *9PM* setup. Here the instruction displayed on the screen will be either the word 'Red' or 'Blue.' The font color of the word may be red or blue. The participant must pick a peg from the area indicated by the font color. For example, if the instruction provided is 'Red' with blue font color, the participant must pick the peg from the blue area. By comparing T2 and T3, we can understand the cognitive effect of the Stroop Test on the participant.

Task 4 (T4) incorporates the WCST into the *9PM* setup. Here, there are two rules; follow the text or follow the color of the text. For example, if the instruction provided is 'Red' with blue font color, the participant must figure out if they need to follow the text and pick from the red area or need to follow the text color and pick from the blue area.

Finally, Task 5 (T5) incorporates the PSMT into the *9PM* setup. Here the participants are shown the order of the pegs in sequence. They are then supposed to memorize this sequence and move the pegs accordingly. They will move the pegs once all nine steps are shown. The participants are asked to do a single round of T1, and 4-rounds of T2, T3, T4, and T5. For more details about the *9PM* tasks and setup, please refer to [1].

3.2 User Study and Data Collection

The user study, which is approved by the Institution's Review Board at the University of Texas at Arlington, has 63 healthy participants. The participants are mostly right-handed (96.82%), male (88.89%) participants, with an average age of 25.11 (± 4.39) years old. Each participant produced 17 datasets (one per round), in addition to the baseline. In T2, T3, T4, and T5, we averaged the rounds of each task after extracting the features to produced another 4 datasets. The data is available at: *https://github.com/abujelala/9PM*.

The participant starts the study by sitting on a chair and watches a video demo[2] which explains the study protocol and how to wear the sensors. At the end of the video demo, the participant reviews and signs the consent form. The study personnel is available in the room to answer any questions the participant might have. After that, the participant wears the sensors. Figure 2 illustrates the experiment setup and the placement of sensors on the participant. The experiment workflow is explained in Figure 3. The sensors record ECG, EDA, EEG, and IMU data. ECG and EDA data are recorded using a Biosignalsplux Explorer unit [6], EEG data is

[2] *https://youtu.be/1O5pmqFOFFQ*

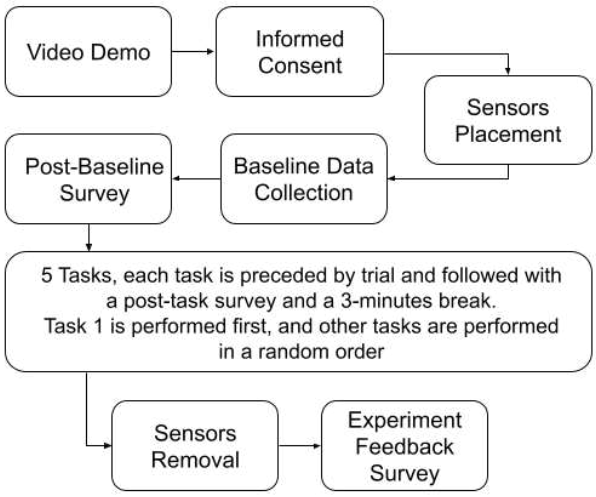

Figure 3: Experiment Workflow.

recorded using an OpenBCI ULTRACORTEX MARK IV sensor [22], and IMU data is recorded using a MetaMotionR sensor [21]. Once the participant wears the sensors and the study personnel checks that the data is reliably streaming, the baseline recording starts. The baseline session is three-minutes long, and it is recorded while the participant is sitting with their eyes closed.

After the baseline, the participant fills out a user questionnaire to collect their baseline subjective data, and then starts performing the study's five tasks. The participants always start with T1 and then perform the other four tasks in random order. Each task is preceded by a trial. The trial of T1 is a shortened version of the task with just 18-moves. However, the trial of T2, T3, T4, and T5 is the same as one round of the task. T1 and the 4^{th} round of T2, T3, T4, and T5 are followed with a lengthy post-task survey while the other rounds are followed with a short post-round survey. The objective of doing T2, T3, T4, and T5 in random order is to focus on the cognitive effect of the tasks and avoid the practice effect and the cascading effect of fatigue. The practice effect is when the participants get very familiar with the setup and perform very well in the last tasks compared to the first tasks. On the other hand, the cascading effect of fatigue is when the performance decreases in the last tasks because the participant is fatigued.

3.3 Machine Learning Approach

3.3.1 Data types and Feature Extraction. In this study, we collected three types of data: behavioral, performance, and physiological data. Behavioral data include the user's dominant-hand movement data and user surveys. The hand movement data is recorded from an IMU sensor attached to the wrist. From the IMU data, we extracted 280 time-domain features based on [2]. These features include the mean, standard deviation, median absolute value, maximum, minimum, signal magnitude area, energy, interquartile range, entropy, autoregression coefficients, and correlation coefficient. The surveys collected the user's subjective responses on mental effort, physical effort, task difficulty, interest in the task, and difficulty concentrating.

The performance metrics in this study are task score, reaction time, and completion time. Task score is the percentage of the correct moves a participant completes. The reaction time is the time the participant needs to decide on which source area they should pick up a peg from. In T2, T3, and T4, the reaction time is calculated from the appearance of a stimulus on the screen to the time a peg is picked up from the source area. In T1 and T5, the reaction time is the time from placing a peg in the destination area to the time a peg is picked up from the source area. In contrast, the completion time is the time the participant needs to complete the physical movement, which is the time needed to place a peg in the destination area after picking it up from the source area.

The physiological data collected are EEG, EDA and ECG. The EEG sensor has 8-channels recording data from the FP1, FP2, P3, P4, C3, C4, O1, and O2 locations on the international 10–20 system for EEG. This was done to monitor the prefrontal and the parietal regions of the brain. These areas have shown to have a relationship with cognitive load and fatigue [9, 38]. Specifically, the parietal cortex has shown a relationship with cognitive fatigue in people suffering from multiple sclerosis. Moreover, EEG was also set up to capture data from the occipital lobe to examine data from the visual cortex. We extract time domain features like mean, standard deviation, minimum, maximum, kurtosis and skew from the EEG data. We also extract spectral and energy based features which have proven to have successful results on such signals [16, 23]. The spectral and energy features are spectral centeroid, spectral spread, spectral entropy, spectral rolloff, zero crossing rate, energy, and entropy of energy. Since these features were extracted for every power band (alpha, beta, gamma, theta and delta) and the raw signal of each EEG channel, the total number of EEG features is 624. From the EDA data, we extract 40 features based on [34]. These features are based on filtered skin conductance signal and its first and second derivatives (e.g. maximum, minimum, maximum of absolute value, and mean absolute value), and wavelet coefficients (maximum, mean, standard deviation, median, and number above zero). For ECG, we extracted 17 time and frequency domain features [7, 31]. These ECG features are mean , median, range, standard deviation, coefficient of variance of NN interval, average heart rate, standard deviation of successive NN interval differences (SDSD), root mean square successive difference (RMSSD), number of pairs of adjacent NN intervals differing by more than 20 ms to all NN intervals and by more than 50 ms (pNN20, pNN50), pNN50/pNN20, power spectrum of low frequency (LF), power spectrum of high frequency (HF), symphathetic modulation index (SMI), vagal modulation index (VMI), and symphatovagal balance index (SVI). The multimodal data used to train the Machine Learning algorithms explained in section 3.3.2 are a combination of these 6 modalities: ECG, EDA, EEG, IMU, performance metrics, and the task number. This leads to 63 possible modality combinations ($2^6 - 1$) for analysis.

3.3.2 Machine Learning Algorithms. We used Machine Learning (ML) to utilize the users' multimodal data to predict their physical and cognitive states. Our goal is to show that data collected from this system can be reliably used to predict cognitive and physical conditions. This paper focuses on predicting 5 states: mental effort, physical effort, task difficulty, task interesting, and difficulty concentrating. The user survey answers were used as the ground truth

of the user's state. Most of the survey questions use a Likert scale from 1 to 10. However, we want to solve the prediction problem as a classification problem, rather than a regression problem. Therefore, we converted the survey answers into a binary scale. For mental effort, physical effort, task difficulty, and task interesting, we try to predict if the user's response is above his/her personal average response (e.g., exerted more than his/her average mental effort). Since we want the prediction to be personalized per participant, each participant's data were normalized separately. For example, if 3 and 9 were the participant's lowest and highest scores, respectively, his/her normalized average score would be 6. Difficulty concentrating was not task-specific, like task difficulty, hence we use ML to predict if the participant is having difficulty concentrating more than in his/her baseline.

As mentioned in section 3.2 and 3.3.1, we have 21 datasets per participant, and every dataset has data from 6 modalities. The 63 possible combinations of the modalities are used to train, validate and test the ML algorithms. The datasets were divided into two parts. The first part has data from 55 participants and it is used for training and validation using 5-fold cross-validation. The second part has 8 participants' data and it is used for testing. The ML algorithms used are Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting (GB), Extra Trees (ET), Decision Tree (DT), Random Forest (RF), Neural Networks (NN), Naive Bayes (NB), AdaBoost (AB), Quadratic Discriminant Analysis (QDA), and Gaussian Process (GP). Randomized Grid-Search from Scikit-Learn Library [30] was used to fine-tune these algorithms.

After the ML features were extracted, we run the ML algorithms multiple times, once with features normalized with Min-Max Scaling, once with features standardized with Standard Scaling, and with and without features selection process. The feature selection process used is Principal Component Analysis (PCA). For every run, we recorded the test F1 score, the test accuracy, and the cross-validation average validation-accuracy. The labels we have are not equally distributed, which makes the accuracy score less reliable. Therefore, the best classifier was determined based on the test F1 score, rather than the test accuracy score. F1 score is a measure of the harmonic mean of the precision and recall [28].It also gives a better understanding of the misclassified cases as it is critical in the design of the framework. We found that the Min-Max normalization with PCA provides the best results for most of the cases. Therefore, section 4 presents the best classifier's test F1 score results when using Min-Max normalization with PCA.

Table 1: ML Results

	Test F1 Score	ML Algorithm
Mental Effort	75.86	QDA
Physical Effort	43.84	NB
Task Difficulty	86.67	AB
Task Interesting	76.92	NB
Difficulty Concentration	55.00	QDA

4 EXPERIMENTAL RESULTS

In this section we focus on the ML algorithms results and the survey results. Table 1 summaries the results of the ML algorithms. From the table, we can see that the ML algorithms result in a high F1 score (> 70%) when predicting mental effort (75.86%), task difficulty (86.67%) and interest in a task (76.92%) but produces a low F1 score when predicting physical effort (43.84%) and difficulty in concentration (55.0%). We believe the reason for the low F1 scores in predicting physical effort and difficulty in concentration is the fact that the participants' responses were not equally distributed. We performed a One-Way Analysis Of Variance (ANOVA) test to analyze user's responses between tasks. We found for most of the tasks there was no significant difference between users' answers on physical effort and difficulty in concentration. Also, when the users' responses were converted to binary scores, the labels were not balanced. The binary labels were 75.28% vs. 24.72 and 72.79% vs. 27.21% for the physical effort and difficulty concentrating, respectively. Since the physical aspect of the tasks is the same, it is expected that the participants might not report changes in their physical effort which justifies the results for the ANOVA test and the low prediction capability for these labels.

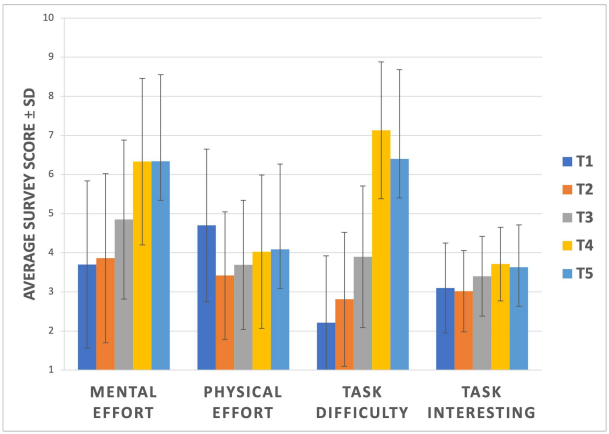

Figure 4: Survey Answers - Average Responses on Mental Effort, Physical Effort, Task Difficulty and Task Interesting.

From the user surveys, we focus on 5 survey questions which we used to produce the ML labels. Figure 4 shows the average survey responses per task with the standard deviation (SD) for mental effort, physical effort, task difficulty and interest in a task. The figure shows noticeable differences in the mean of the mental effort and the task difficulty responses while it shows small mean differences in responses of the other two questions. The figure also shows that T4 and T5 required the most mental effort and were the most difficult. It also shows that T1 required the most physical effort. The average interest in a task was also very similar across the tasks, with T4 being the most interesting task. In addition, Figure 5 illustrates the average responses for difficulty concentrating. It shows that the most common response was *'No More than Usual'*.

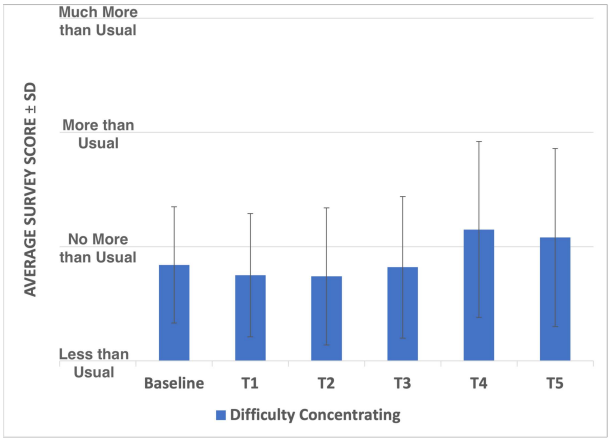

Figure 5: Survey Answers - Average Responses on Difficulty Concentrating.

5 CONCLUSION

In this study, we developed the *9PM* system, a novel system to bridge the gap between real-world activity and clinical tests. *9PM* was designed to utilize on the standardized physical and cognitive assessment and to collect multimodal data for analysis and assessment. We were able to analyze these multimodal data and use them to predict user's state. In this paper, our analysis mainly focuses on ML and survey responses. Results indicate that the data collected using the system can be used to reliably predict cognitive state. The survey responses also indicate that the *9PM* system can be used to build cognitive tests of varying difficulty and a physical task using the same system. The future goal of our research is to build an adaptive intelligent system that could track user's state and provide personalized recommendations. To do that, we need to detect the user's current state, and we believe our results in this paper provide the necessary motivation to use the *9PM* system towards our goal. While our work in this paper provides a successful proof of concept. The data in this paper was collected in a controlled lab environment, with healthy participants in the same age range. Therefore, our system still needs to be tested and validated with more diverse participants, in a real-life scenario.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grant No. 1719031. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Maher Abujelala et al. 2019. *THINK2ACT: USING MULTIMODAL DATA TO ASSESS HUMAN COGNITIVE AND PHYSICAL PERFORMANCE.* Ph.D. Dissertation.

[2] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones.. In *Esann*, Vol. 3. 3.

[3] Ashwin Ramesh Babu, Akilesh Rajavenkatanarayanan, James Robert Brady, and Fillia Makedon. 2018. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signal. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. 1–7.

[4] Francisco Barceló, Juan M Muñoz-Céspedes, Miguel A Pozo, and Francisco J Rubia. 2000. Attentional set shifting modulates the target P3b response in the Wisconsin card sorting test. *Neuropsychologia* 38, 10 (2000), 1342–1355.

[5] Patricia J Bauer, Sureyya S Dikmen, Robert K Heaton, Dan Mungas, Jerry Slotkin, and Jennifer L Beaumont. 2013. III. NIH Toolbox Cognition Battery (CB): measuring episodic memory. *Monographs of the Society for Research in Child Development* 78, 4 (2013), 34–48.

[6] Biosignalsplux [n.d.]. Biosignalsplux | wearable body sensing platform - Explorer. https://www.biosignalsplux.com/en/explorer

[7] Sansanee Boonnithi and Sukanya Phongsuphap. 2011. Comparison of heart rate variability measures for mental stress detection. In *2011 Computing in Cardiology*. IEEE, 85–88.

[8] Noelle E Carlozzi, Jennifer L Beaumont, David S Tulsky, and Richard C Gershon. 2015. The NIH toolbox pattern comparison processing speed test: normative data. *Archives of Clinical Neuropsychology* 30, 5 (2015), 359–368.

[9] John DeLuca, Helen M Genova, Frank G Hillary, and Glenn Wylie. 2008. Neural correlates of cognitive fatigue in multiple sclerosis using functional MRI. *Journal of the neurological sciences* 270, 1-2 (2008), 28–39.

[10] Sureyya S Dikmen, Patricia J Bauer, Sandra Weintraub, Dan Mungas, Jerry Slotkin, Jennifer L Beaumont, Richard Gershon, Nancy R Temkin, and Robert K Heaton. 2014. Measuring episodic memory across the lifespan: NIH toolbox picture sequence memory test. *Journal of the International Neuropsychological Society* 20, 6 (2014), 611–619.

[11] Gammon M Earhart, Jim T Cavanaugh, Terry Ellis, Matt P Ford, K Bo Foreman, and Lee Dibble. 2011. The 9-hole PEG test of upper extremity function: average values, test-retest reliability, and factors contributing to performance in people with Parkinson disease. *Journal of Neurologic Physical Therapy* 35, 4 (2011), 157–163.

[12] Peter Feys, Ilse Lamers, Gordon Francis, Ralph Benedict, Glenn Phillips, Nicholas LaRocca, Lynn D Hudson, Richard Rudick, and Multiple Sclerosis Outcome Assessments Consortium. 2017. The Nine-Hole Peg Test as a manual dexterity performance measure for multiple sclerosis. *Multiple Sclerosis Journal* 23, 5 (2017), 711–720.

[13] James M Fisher, Nils Y Hammerla, Lynn Rochester, Peter Andras, and Richard W Walker. 2016. Body-worn sensors in Parkinson's disease: Evaluating their acceptability to patients. *Telemedicine and e-Health* 22, 1 (2016), 63–69.

[14] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 39–42.

[15] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. 2019. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing* (2019).

[16] Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one* 10, 12 (2015), e0144610.

[17] David A Grant and Esta Berg. 1948. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of experimental psychology* 38, 4 (1948), 404.

[18] Ryan Hays, Philip Henson, Hannah Wisniewski, Victoria Hendel, Aditya Vaidyam, and John Torous. 2019. Assessing cognition outside of the clinic: smartphones and sensors for cognitive assessment across diverse psychiatric disorders. *Psychiatric Clinics* 42, 4 (2019), 611–625.

[19] Varun Kanal, James Brady, Harish Nambiappan, Maria Kyrarini, Glenn Wylie, and Fillia Makedon. 2020. Towards a serious game based human-robot framework for fatigue assessment. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–6.

[20] Virgil Mathiowetz, Gloria Volland, Nancy Kashman, and Karen Weber. 1985. Adult norms for the Box and Block Test of manual dexterity. *American Journal of Occupational Therapy* 39, 6 (1985), 386–391.

[21] MetaMotionR - MbientLab [n.d.]. MetaMotionR - MbientLab. https://mbientlab.com/metamotionr/

[22] OpenBCI [n.d.]. Ultracortex Mark IV | OpenBCI Documentation. http://docs.openbci.com/Headware/01-Ultracortex-Mark-IV

[23] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2019. CogBeacon: A Multi-Modal Dataset and Data-Collection Platform for Modeling Cognitive Fatigue. *Technologies* 7, 2 (2019), 46.

[24] Janet L Poole, Patricia A Burtner, Theresa A Torres, Cheryl Kirk McMullen, Amy Markham, Michelle Lee Marcum, Jennifer Bradley Anderson, and Clifford Qualls. 2005. Measuring dexterity in children using the Nine-hole Peg Test. *Journal of Hand Therapy* 18, 3 (2005), 348–351.

[25] Sen Qiu, Zhelong Wang, Hongyu Zhao, Long Liu, and Yongmei Jiang. 2018. Using body-worn sensors for preliminary rehabilitation assessment in stroke victims with gait impairment. *IEEE Access* 6 (2018), 31249–31258.

[26] Akilesh Rajavenkatanarayanan, Varun Kanal, Konstantinos Tsiakas, James Brady, Diane Calderon, Glenn Wylie, and Fillia Makedon. 2019. Towards a robot-based

multimodal framework to assess the impact of fatigue on user behavior and performance: a pilot study. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 493–498.

[27] A. Rajavenkatanarayanan, H. R. Nambiappan, M. Kyrarini, and F. Makedon. 2020. Towards a Real-Time Cognitive Load Assessment System for Industrial Human-Robot Cooperation. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 698–705. https://doi.org/10.1109/RO-MAN47096.2020.9223531

[28] Yutaka Sasaki. 2007. The truth oh the F–measure. *Manchester: School of Computer Science, University of Manchester* (2007).

[29] Federica Scarpina and Sofia Tagini. 2017. The Stroop Color and Word Test. *Frontiers in Psychology* 8 (2017), 557. https://doi.org/10.3389/fpsyg.2017.00557

[30] scikit [n.d.]. Scikit-Learn: RandomizedSearchCV. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

[31] Fred Shaffer and JP Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* 5 (2017), 258.

[32] J Ridley Stroop. 1935. Studies of interference in serial verbal reactions. *Journal of experimental psychology* 18, 6 (1935), 643.

[33] Jeanette Taylor. 2007. Heritability of Wisconsin Card Sorting Test (WCST) and Stroop Color-Word Test performance in normal individuals: implications for the search for endophenotypes. *Twin Research and Human Genetics* 10, 6 (2007), 829–834.

[34] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. 2015. Automatic identification of artifacts in electrodermal activity data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1934–1937.

[35] Konstantinos Tsiakas, Maher Abujelala, and Fillia Makedon. 2018. Task engagement as personalization feedback for socially-assistive robots and cognitive training. *Technologies* 6, 2 (2018), 49.

[36] David S Tulsky, Noelle Carlozzi, Nancy D Chiaravalloti, Jennifer L Beaumont, Pamela A Kisala, Dan Mungas, Kevin Conway, and Richard Gershon. 2014. NIH Toolbox Cognition Battery (NIHTB-CB): The list sorting test to measure working memory. *Journal of the International Neuropsychological Society: JINS* 20, 6 (2014), 599.

[37] Simone Varrasi, Santo Di Nuovo, Daniela Conti, and Alessandro Di Nuovo. 2018. A social robot for cognitive assessment. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 269–270.

[38] Robert R Whelan. 2007. Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review* 2, 1 (2007), 1–12.