Enhancing User Engagement in DailyExp: A Tool for Collecting Cognitive Performance and Physiological Data with Engaging Behavioral Design

Xianyin Hu
Graduated School of Frontier Sciences
The University of Tokyo
Chiba, Japan
Email: huxy214@gmail.com

Yuki Ban
Graduated School of Frontier Sciences
The University of Tokyo
Chiba, Japan
Email: ban@edu.k.u-tokyo.ac.jp

Shin'ichi Warisawa
Graduated School of Frontier Sciences
The University of Tokyo
Chiba, Japan
Email: warisawa@edu.k.u-tokyo.ac.jp

***Abstract*—Experiments in cognitive science that rely on laboratory-based settings are not only costly and time-consuming but also make it difficult to investigate individuals' cognitive states as they naturally fluctuate over time in daily life. In our initial study [1], we presented a practical tool implemented as a smartphone application DailyExp that aims to conveniently collect cognitive performance data in daily life settings. This application is accessible from major mobile platforms (iOS and Android), tied with a Fitbit account to collect physiological data at the same time. We employed engaging behavioral design to overcome problems faced experimenting in the wild, intended to improve data quality as well as data collection efficiency and evaluated them in a one-month-long experiment involving 10 participants. For this extended study, we implemented new features based on feedback from the preliminary study and tested them on 41 individuals. This paper provides implementation details of each cognitive task that was not covered in the initial paper, and included a comprehensive analysis of post-study questionnaires as well as a quantitative comparison of objective metrics evaluating user engagement and persistence. Our results demonstrated a statistically significant improvement in user engagement as well as persistence by adding new features.**

***Index Terms*—*data collecting tool*; *engaging design*; *physiological data*; *cognitive performance*;**

I. Introduction

Cognitive science has traditionally centered on comprehending the mechanisms of human cognition at an aggregate level. However, the exploration of individual differences has emerged as a progressively significant subject within the field. Recently, researchers have adopted a perspective that views individuals' cognition as a dynamic system that fluctuates. It is also suggested that the fluctuation in cognition is related to fluctuation in physiology from an embodied cognition perspective [2]. In this study, Our primary interest lies in facilitating studies that investigate individual differences and intra-individual variations within established cognitive mechanisms. We advocate for these investigations to be carried out in meticulously designed real-life settings, as opposed to laboratory settings, since laboratory environments may induce high arousal levels that shift individuals' cognitive and physiological states due to nervousness and unfamiliarity. Experiments carried out with authentic context ensure the accumulation of extensive data encompassing both population-wide variations and intrapersonal dynamics. To support such endeavors, we offered a practical tool DailyExp tailored to assist researchers in gathering data on cognitive performance and physiological signals within daily life settings.

Previous attempts have been made to adopt smartphones and smartwatches as assessment tools, including iVitality [3], DelApp [4], Cognition Kit [5] and UbiCAT [6]. These studies showed a good correspondence between data obtained from the mobile-based tools and that from the laboratory, indicating that mobile-based tools are feasible for evaluating cognitive function. However, challenges such as a lack of user engagement throughout a prolonged experiment still exist in an experiment conducted in the wild that depends largely on participants' voluntary behaviors.

This study provided a practical implementation of a tool for data collection of both cognitive performance, as well as physiological data in daily life settings with engaging behavioral design. An alpha version of the smartphone application DailyExp is readily available that can conduct various classical paradigms in cognitive science including N-back, Stroop, and Raven's Advanced Progressive Matrices (RAPM) test. The application was linked with a widely used commercial smartwatch (Fitbit) to collect physiological data at the same time. This application is accessible from major mobile platforms (iOS and Android). We employed multiple practices of engaging behavioral designs to overcome several challenges facing experimenting in the wild, including immediate reward and feedback, trackable performance to bolster user self-efficacy, transparent monetary rewards for enhanced psychological safety, and a ranking system to ignite social competition.

Building upon the features previously evaluated with 10 participants in our initial conference paper [1], we made two key modifications to DailyExp and tested on 41 individuals to address user feedback collected in the preliminary study. Firstly, we revised the ranking screen to display only the top 15 most active participants, concealing information about less-engaged users to prevent demotivation among others. Secondly, we implemented a daily limit on task completion to foster a sense of achievable goals and discourage procrastination. This adjustment was based on user comments highlighting the negative impact of unrestricted task completion on motivation. Also, this paper provided implementation details of each cognitive task that was not covered in the initial paper, and included a comprehensive analysis of post-study questionnaires as well as a quantitative comparison of objective metrics evaluating user engagement and persistence. Our results demonstrated a statistically significant improvement in

user engagement as well as persistence by adding new features.

The paper is structured as follows: In Section II, we reviewed existing literature on the utilization of smartphones and smartwatches as assessment tools, explored the feasibility of using mobile devices for data collection in daily life and highlighted the challenges associated with this approach. In Section III, we provide a detailed description of the implementation that covers the technical aspects of developing the tool. In Section IV, we evaluated the system through a preliminary user study involving ten participants over one month. We compiled several points for improvement based on the user interview in the preliminary study and subsequently enhanced the DailyExp. In Section V, we recruited 41 participants to use the improved version of DailyExp for one month. We compared the objective metrics including user engagement and persistence with those in the preliminary study and revealed the effectiveness of the enhancement. Both experiments in this study were approved by the Research Ethics Committee of the University of Tokyo (No. 22-399) and performed under written informed consent from all the participants. Finally, we present a summary in Section VI that highlights the achievements of this paper and discusses the potential applications of DailyExp in different domains of cognitive science, showcasing its versatility.

II. Related Works

In this section, we reviewed existing literature on the utilization of mobile devices as assessment tools, explored the feasibility of using mobile devices for data collection in daily life and highlighted the challenges associated with this approach.

Jongstra's team [3] developed a smartphone-based app iVitality to evaluate five cognitive tests (Memory-Word, Trail Making, Stroop, Reaction Time, and Letter N-back) in 151 healthy adults over 6 months. This study concluded that repeated smartphone-assisted cognitive testing is feasible with reasonable adherence and moderate validity for the Stroop and the Trail Making tests compared with conventional neuropsychological tests. They also addressed that smartphone-based cognitive testing seems promising for large-scale data collection in population studies.

Tieges' team developed the DelApp [4], a smartphone application aimed at objectively detecting attentional deficits in delirium patients. By leveraging mobile technology, this study addressed the need for accessible and accurate cognitive assessment tools in clinical settings, contributing to improved diagnosis and management of delirium. Moreover, this study showed the use of smartphone-based cognitive tasks allows for a broader range of participants, extending beyond healthy individuals to include those with varying cognitive conditions.

Dingler introduced a mobile toolkit designed to capture fluctuations in alertness and cognitive performance throughout the day [5]. The tool kit consists of three tasks, they are Psychomotoric Vigilance Task, Go/No-Go task, and Multiple Object Tracking task. To investigate the feasibility of using smartphone-based cognitive tasks to assess diurnal fluctuations in arousal level, the researchers conducted an in-the-wild experiment with 12 participants who completed the test batteries multiple times a day over 1-2 weeks. The results demonstrated that circadian rhythms in alertness could be effectively captured in naturalistic settings, without the need for controlled laboratory conditions. This study provided insights into the utilization of smartphone-based cognitive test batteries to capture temporal dynamics of cognitive function, showing their potential applications in personalized healthcare and productivity management.

Hafiz [6] presented the UbiCAT, a smartwatch-based cognitive assessment tool. This tool implemented 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. They evaluated the UbiCAT test measures against standard computer-based tests with 21 healthy adults. The results showed a strong correlation between the UbiCAT and standard computer-based tests, indicating its effectiveness. Usability ratings were high, and participants reported low discomfort while using the smartwatch. Despite some participants preferring computer-based tests due to familiarity, the UbiCAT offered the advantage of in-the-wild assessment, leveraging the ubiquity of wearable devices. However, it is limited by the less diverse interaction methods available on smartwatches compared to smartphones.

In summary, these studies demonstrated the feasibility of leveraging mobile devices for collecting cognitive performance data in a real-world setting. While smartwatches provide convenient assessment, smartphones offer a wider range of interaction, making them suitable for assessing a large variety of cognitive functions. However, a common challenge of these studies is maintaining participant engagement, as the amount and quality of data depend largely on voluntary participation. In this study, we provided a practical implementation of a tool for collecting not only cognitive performance data but also physiological signals linked with the widely used commercial smartwatch Fitbit. These physiological signals can be used to predict cognitive function and explore the underlying mechanisms of cognition and physiology. Moreover, instead of validating how well mobile-based cognitive tests align with results from laboratory settings, which is done in the precedent studies, we addressed the practice of engaging behavior design, aiming to enhance the quality and consistency of data collection in naturalistic settings.

III. Implementation

This section covered the technical aspects of developing the DailyExp. First, we provide a systematic overview. Then we described the implementation of the three cognitive tasks-Spatial 2-back, Stroop, and FluidIQ in detail, which was not covered in the initial conference paper due to page constraints. Finally, we demonstrated the features designed to handle unexpected user behavior and those implementing best practices for engaging behavioral design.

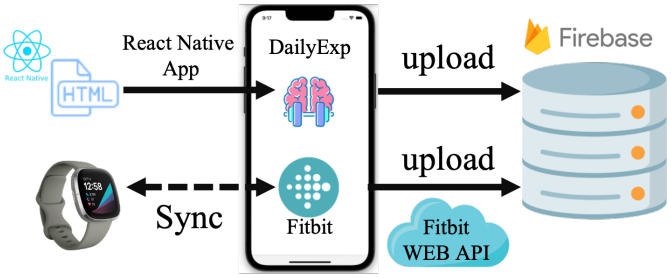
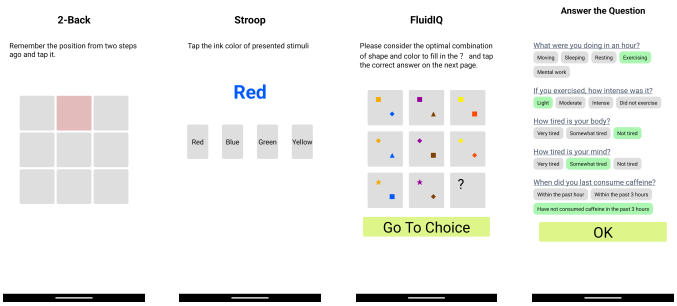


Fig. 1: Overview of the system design.



(a) (b) (c) (d)

Fig. 2: Screenshots of DailyExp. (a) The 2-back task. (b) The Stroop task. (c) The FluidIQ task. (d) The post-task questionnaire.

A. System Overview

Fig. 1 showed the overview of the system design. The mobile client of DailyExp was developed using React Native, a web-based open-source framework for mobile application development. React Native was chosen to ensure compatibility across multiple platforms for iOS and Android devices. For the server side, Firebase's data storage service was utilized to store data, including users' daily summary data, cognitive performance data for various tasks, and physiological data grabbed from the Fitbit server using Fitbit web API.

B. Cognitive Tasks

In the alpha version of DailyExp, three well-established cognitive tasks were administrated to study working memory (N-back), attention and executive function (Stroop), and fluid intelligence (FluidIQ or Raven's Progressive Matrices) as shown in Fig. 2. (a)-(c). These tasks were selected due to their robustness and potential to have individual differences and intrapersonal fluctuations in the corresponding cognitive ability to be evaluated. Cognitive Performance data with the trial information, problem context, and users' responses will be recorded and uploaded to the Firebase server. The details of the implementation of the three tasks were described as follows.

1) The Spatial Continuous 2-back Task: The n-back task paradigm is commonly used as an assessment in psychology and cognitive neuroscience to measure a part of working memory and working memory capacity. In this study, we employed a variation of the N-back task called the continuous N-back task. As suggested in previous work [7], the performance of continuous 2-back tasks reflected the level of mental fatigue and can be estimated from physiological signals using a deep learning model. Since the purpose of this study is to collect data that can be used to investigate intra-day fluctuations in individuals' cognitive states, we considered the continuous version instead of the original one a better candidate.

In the continuous N-back task, instead of responding only at the matched trials (when the current stimulus matches the stimulus presented in N trials before), but continuously respond with the stimulus presented in N trials before. Regarding the type of stimuli, we adopted spatial position, which aligns best with the touch-based input method on smartphones. The stimuli sequence was randomly generated. The task lasted for 5 minutes, and the stimuli were presented every 2 seconds. Thus there will be a total of 150 trials in each conduction.

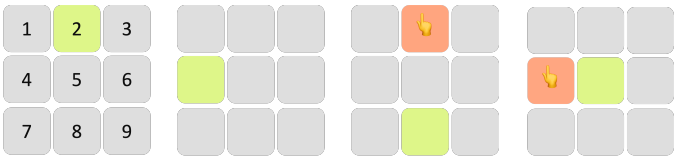
Fig. 3 illustrated an example of four trials in the spatial continuous 2-back task. For explanatory purposes, numbers are assigned to each position in Fig. 3a, although these numbers were not displayed in the actual task. The presented position was highlighted in green to distinguish it from other positions. Participants were required to respond by tapping the corresponding position on the smartphone screen. The tapped position turned red to provide feedback. In this example, a sequence of positions (2, 4, 8, 5) were presented. On the third trial, when position 8 was presented, a correct response would be tapping position 2, which had been presented two trials before. Similarly, on the fourth trial, it was desired to respond by tapping position 4. Additionally, to assist users in developing an intuitive sense of the presentation intervals, the green color highlighting the current stimulus gradually faded over the 2-second interval.

Cognitive performance data collected in the spatial continuous 2-back task were listed in Table I. Specifically, we recorded stimulus up to three trials back to facilitate the analysis of correctness as well as error patterns. Regarding user response, both the response position and reaction time were recorded to enable a multidimensional analysis of human cognitive behavior.

2) The Stroop Task: The Stroop task was originally devised by John Ridley Stroop in 1935 to investigate the interference in serial verbal reactions [8], has been widely used in cognitive science and experimental psychology to study selective attention [9]. In the Stroop task, participants are presented with

TABLE I: Cognitive performance data recorded in spatial continuous 2-back task

Trial infomation	trial ID trial start time
Problem context	current stumilus 1-back stimulus 2-back stimulus 3-back stimulus
User response	responded position reaction time



(a) 1st trial (b) 2nd trial (c) 3rd trial (d) 4th trial

Fig. 3: 2-back detail explained.2-back detail explained.2-back detail explained.2-back detail explained.2-back detail explained

TABLE II: Cognitive performance data recorded in Stroop task

Trial infomation	trial ID trial start time
Problem context	stimulus word stimulus ink color
User response	responded color reaction time

words printed in specific colors. In the congruent condition, the color of the word matches the meaning of the word. In the incongruent condition, they do not match. Participants are asked to either say aloud the word or to report the color of the word. As a control condition, participants read words printed in black or report the color of non-word stimuli.

The Stroop task was implemented in DailyExp as shown in Fig. 2b. As mentioned in the previous section, this study aims to investigate the fluctuations in individual cognitive states, rather than the already validated Stroop effect. Therefore, we only presented the incongruent condition where the ink color of the stimulus was different from the meaning of the word, leading to a fairly difficult level for most of the users.

In each trial, two different colors were randomly selected to be presented as word or ink colors from a set of four colors: red (#ff4b00), blue (#005aff), green (#03af7a), and yellow (#fff100). Those colors are recommended in Color Universal Design, making them accessible to people with diverse color vision characteristics. Participants were instructed to tap the button corresponding to the ink color of the presented stimulus, and the presentation lasted for 3 seconds. To indicate there is a transition to the next trial, a 2-second retention interval was included after each presentation. During the retention interval, nothing was displayed. Consequently, the interval for each trial was 5 seconds. Cognitive performance data collected in the Stroop task were listed in Table II.

3) The FluidIQ Task: The RPM (Raven's Progressive Matrices) task, also known as the FluidIQ task, presents an incomplete geometric pattern and requires participants to select the missing piece from a set of options. The RPM task is a non-verbal test that does not require any specific knowledge or language skills. It is used worldwide to measure general human intelligence and abstract reasoning abilities and gained recognition as an IQ test since it was adopted as a part of the Monsa intelligence assessment. Intelligence is generally classified into two categories which are crystallized and fluid intelligence. Crystallized intelligence represents individual abilities acquired over years of experience, education, and learning, while fluid intelligence represents insight and abilities independent of language or experiential knowledge. The non-verbal nature and the independence from specific knowledge of the task make it suitable for evaluating fluid intelligence. For this reason, it is also referred to as the FluidIQ task. Since the latter term more intuitively captures the essence, we will adopt the term FluidIQ in this study.

The FluidIQ task used in this study was implemented as shown in Fig. 2c, and a set of features as described in Table III were generated for each trial. The generation of the whole figure followed a certain set of rules. There are two possible rules or different-along the row and the column. Rules were randomly assigned to row and column despite a case when rule same was assigned to both since it would make the task too easy to be solved intuitively without involving the desired cognitive process. In summary, there were three possible combinations of row and column rules:

- Same along the row, different along the column.
- Different along the row, same along the column.
- Different along the row, different along the column.

If all four features follow the rule different along row/column and the rule same along the other axis, the total number of features needed to be generated would be 4, also denoted as $n^{\text{different}} = 4$. If all four features follow the rule different along both row and column, then $n^{\text{different}} = 8$, which is the most difficult. For example, in Fig. 2c, for Feature 1 (shape of the top-left figure), the rule along the row is same, and rule along the column is different. For Feature 2 (color of the top-left figure), the rule along the row is different, and rule along the column is same. For Feature 3 (shape of the bottom-right figure), both row and column rules are different. For Feature 4 (color of the bottom-right figure), the row pattern is different, and the column pattern is the same. In summary, for this trial, the number of different rules is $n^{\text{different}} = 5$. In this task, the systematic factor that determines the difficulty level is the number of different rules $n^{\text{different}}$, with possible values of $n^{\text{different}} \in [4, 5, 6, 7, 8]$.

To control individual execution strategies, we implemented the FluidIQ task to not display the problem and options at the same time. Instead, participants were required to observe an incomplete figure to distinguish the rules in the problem view and generate their answers. Participants then tap the "Go to Choice" button below to move to the option view. Once moved on to the option view, participants could not go back to the problem view. This implementation prevented participants from using a strategy of applying each option to the incomplete figure one by one, thereby achieved to force users to adopt a generative strategy.

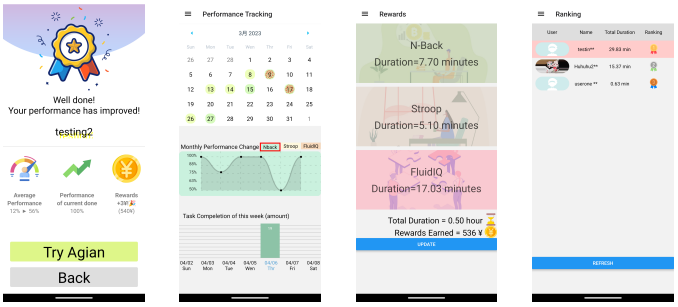
There was no specific time limit for this task, and participants were allowed to perform at their own pace. The FluidIQ task consisted of 12 trials, taking around 5 minutes to complete. The cognitive performance data recorded was shown in Table IV.

TABLE III: Features and corresponding possible values in the FluidIQ task

Feature	Description	Possible values
feature1	shape of the top-left figure	circle, star, triangle, diamond, square
feature2	color of the top-left figure	red(#ff4b00), blue(#005aff), green(#03af7a), orange(#f6aa00), purple(#990099), brown(#804000), yellow(#fff100)
feature3	shape of the bottom-right figure	same as feature1
feature4	color of the bottom-right figure	same as feature2

TABLE IV: Cognitive performance data recorded in FluidIQ task

Trial infomation	trial ID trial start timestamp
Problem context	stimuli combination in position 1 stimuli combination in position 2 ... stimuli combination in position 8 stimuli combination in answer stimuli combination in alternative choice1 stimuli combination in alternative choice2
User response	responded stimuli combination reaction time



(a) (b) (c) (d)

Fig. 4: Screenshots of DailyExp. (a) The encouragement view. (b) The performance tracking view. (c) The rewards view. (d) The ranking view.

C. Dealing with Unexpected User Behavior

An issue facing the experiment that relies on users' voluntary behavior is that users do not always behave in a desired manner. Predictable behaviors include forgetting to wear the smartwatch, responding randomly without involving the target cognitive process to be assessed, and an improper understanding of the procedures for the tasks. These behaviors will lead to a lack of data and noisy meaningless data. To address these issues, our application implemented several features to reduce unexpected behaviors.

Firstly, we provided reminders to wear the smartwatch before the start of any task. Secondly, practice mode with feedback will help users familiarize themselves with the task procedure. Moreover, user performance is recorded to monitor if it falls below a preset threshold. Users will be notified of invalid conduction due to their suboptimal performance.

Another concern facing an experiment in the wild is that it is difficult to control factors that are not the target of interest, such as physical activities and caffeine intake, which have a great influence on the cognitive and physiological state. As a solution, we provide a self-report questionnaire (Fig. 2. (d)) after completing the task. This enables data to be nicely categorized and analyzed afterward.

D. Engaging Behavioral Design

We leveraged multiple practices of engaging behavioral design aiming to improve the efficiency of data collection, which is largely determined by user engagement. Fig. 4. (a) showed the encouragement view that popped up immediately after each task conduction, notifying users about how well they performed this time compared to the past. We expect this action-reward link to lead to the habitation of voluntary conduction of cognitive tasks. Fig. 4. (b) showed a performance tracking view from a long-term perspective. The calendar and line chart displayed the monthly task executions and performance fluctuations, while the bar chart showed the task executions for the current week. This feature took advantage of human's tendency to make more effort towards specific goals when they feel in control of their actions. We expect this feature to satisfy users' desire for autonomy and increase intrinsic motivation. Fig. 4. (c) illustrated the monetary rewards earned, along with detailed information, such as the amount of time spent on each task and the corresponding rewards. In addition to providing an external motivation, this feature also promotes transparency of the experiment and is expected to enhance the psychological safety of participants. Finally, a ranking view (Fig. 4. (d)) was implemented as a social motivation leveraging the competitive mindset by allowing users to see others' task executions.

IV. Preliminary User Study

To evaluate the effectiveness of collecting data of the alpha version of DailyExp as well as gain insights into application design, we conducted a preliminary study involving 10 users to use DailyExp in their daily life for one month. Both qualitative and quantitative analysis was performed.

A. Participants

We recruited 10 participants (6 males and 4 females, aged 21-27) as app users, who are graduated students from the University of Tokyo.

B. Procedure

All participants were scheduled to attend an orientation in our lab. At the beginning of the orientation, we explained the purpose and the procedure of the experiment. After that, the participants signed with informed consent if they agreed to participate in the experiment. Then, they were guided to install Fitbit and DailyExp on the smartphones that they use in daily life. The Fitbit application was necessary to synchronize with the Fitbit device for physiological data collection. The researcher distributed pre-assigned Fitbit accounts to the

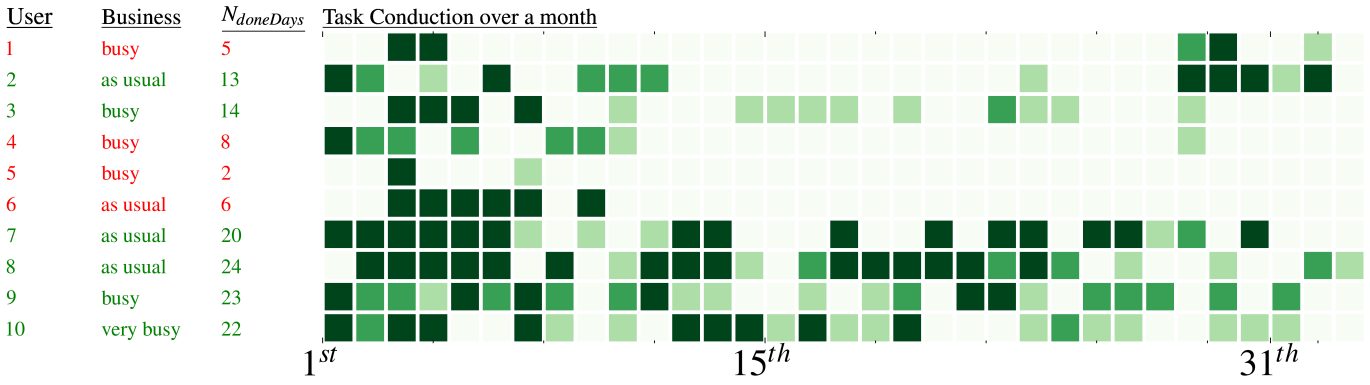


Fig. 5: User's busyness and task conduction in the preliminary user study. $N_{doneDays}$ denotes the number of days with task conducted. The cell color gradient indicated the number of task types performed (etc., the darkest grey indicated a completion of all three different tasks). Users printed in green ink are those who conducted tasks for more than 10 days and were considered active.

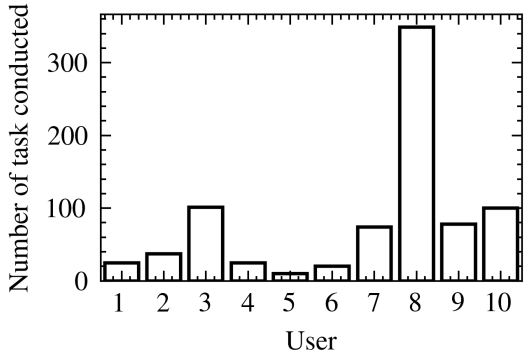


Fig. 6: Number of tasks conducted by each user in the preliminary user study.

participants and assisted them in logging in to DailyExp using Fitbit account. Subsequently, a document containing screenshots (Fig. 2) and descriptions of how to use DailyExp was presented. The participants were allowed to practice the cognitive tasks multiple times until they fully understood how to perform the three tasks correctly. Participants were informed that a monetary reward of 100 Japanese yen was given for each valid task completion. Besides the basic usage of conducting cognitive tasks, the researcher also provided explanations on other functionalities of DailyExp (Fig. 4) to the participants, and have them give a try on them. It has to be addressed that, participants were informed to conduct the cognitive tasks and use other functionality within the DailyExp at their own pace, without any restriction on the number of tasks to be completed each day or any limitations regarding task completion. Finally, instructions about the usage of the Fitbit device, the Fitbit application, and the synchronization process between them were provided. All the participants were requested to return the Fitbit devices at the end of the experiment and fill out a brief questionnaire. The questions and options for answers were listed in Table V.

C. Quantitative Evaluation

1) Data Collected: Throughout the preliminary user study, we obtained a total of 847 rounds of cognitive performance data, with 235 rounds for the spatial continuous 2-back task, 290 rounds for the Stroop task, and 322 rounds for the FluidIQ task.

2) User Engagement: We evaluated user engagement using two metrics. The first metric was the ratio of active users. A user is considered active if he/she conducted task for more than ten days throughout one month. Fig. 5 visualized the ten participants' task completion during the experiment period. It highlighted active users in green and non-active users in red, with their number of days when task was conducted at the left, denoted as $N_{doneDay}$, along with their business reported in the after-study questionnaire. As a result, four out of the ten users (users 7, 8, 9, and 10) actively engaged with DailyExp. Consequently, the ratio of active users in the preliminary user study was 60%. Notably, among the four non-active users, three reported themselves being busier than usual.

The other metric was the number of tasks conducted per user. As shown in Fig. 6, User 8 conducted 349 tasks and contributed to almost half of the total completions. Therefore, we treated User 8 as an outlier and calculated the remaining users' averaged completions, which is around 55 per user.

D. Qualitative Evaluation

1) Reason of losing motivation: In the preliminary user study, the top reasons cited for failure to sustain task execution were being busy during the period (50%) and tasks lasting too long (40%).

2) Functionality that increases motivation: Besides monetary rewards, The Ranking screen was highly rated as contributing to increased motivation for task execution (90%).

3) Possible improvements: The top suggested improvements were shorter task duration (80%), push notifications as reminders (60%) as well as more interesting tasks (60%).

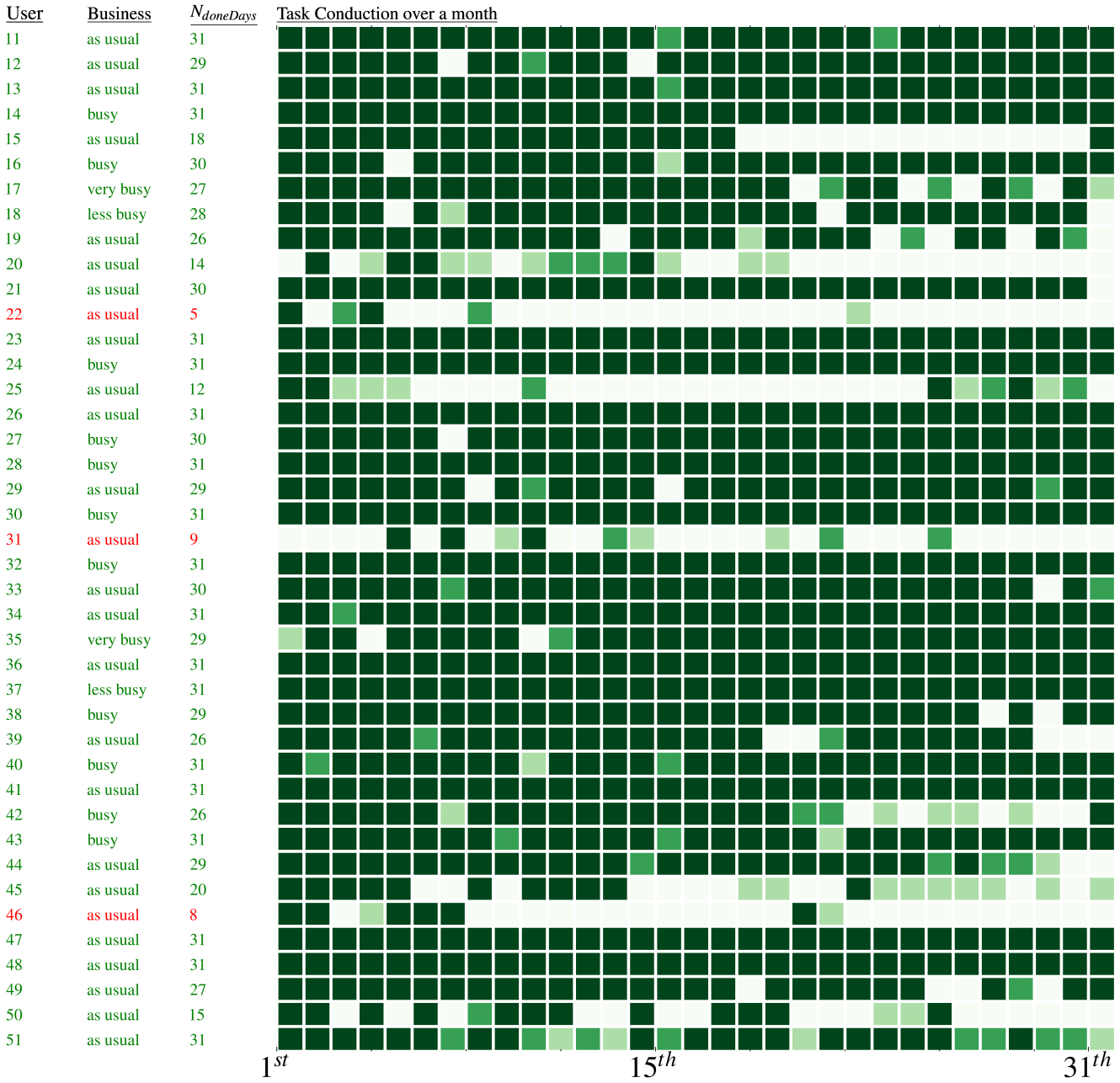


Fig. 7: User's busyness and task conduction in the user study. $N_{doneDays}$ denotes the number of days with task conducted. The cell color gradient indicated the number of task types performed (etc., the darkest grey indicated a completion of all three different tasks). Users printed in green ink are those who conducted tasks for more than 10 days and were considered active.

4) Free comments from users: Two users provided comments about a disadvantage of DailyExp to sustaining motivation as follows:

> User 1: Since all users' task completion status can be seen on the ranking screen, I have a feeling that it was okay not to rush, after observing those who were not executing tasks much. This led to a decrease in my motivation for task execution.

> User 9: My motivation was to earn rewards, and since there are no restrictions on task execution each day, I believed a large amount of tasks could be completed in the last few days, so I procrastinated.

V. User Study with Improved DailyExp

To address the issues identified in the preliminary user study conducted with the alpha version of DailyExp, the following changes were made in the improved version:

TABLE V: Summary of Participant Responses in Post-Study Questionnaire. The last two columns indicated the number and percentage of participants who chose each answer option in the preliminary user study and user study using improved DailyExp, respectively. The top two most chosen answers for each question are highlighted in bold.

question	answer	Preliminary user study	User study
During the experiment period, did your research or work become busier than you usually are?	not busy at all	0 (0%)	0 (0%)
	less busy	0 (0%)	2 (5%)
	as usual	**4 (40%)**	**26 (63%)**
	busy	**5 (50%)**	**11 (27%)**
	very busy	1 (10%)	2 (5%)
Please select the possible reason(s) listed below that contribute to the failure of sustaining task execution. (multiple choices possible)	**I persisted**	**6 (60%)**	**35 (85%)**
	I am busy during the period	**5 (50%)**	3 (7%)
	The tasks were too boring	3 (30%)	0 (0%)
	The tasks last too long	4 (40%)	3 (7%)
	It was bothersome to wear Fitbit device in daily life	2 (20%)	0 (0%)
	It is bothersome to answer the after task questionnaire	0 (0%)	0 (0%)
	I forget to open the app	3 (30%)	**5 (12%)**
From the following functionality of DailyExp, please select the one(s) that contributed to increasing your motivation for task execution. (multiple choices possible)	Performance Tracking	2 (20%)	6 (15%)
	Reward	**7 (70%)**	**21 (51%)**
	Ranking	**9 (90%)**	**32 (78%)**
	Encouragement right after task conduction	2 (20%)	15 (37%)
From the following possible improvements, please select the one(s) you believe would enhance the engagement and persistence for your task execution. (multiple choices possible)	Push notifications to remind me	**6 (60%)**	10 (24%)
	more monetary rewards	4 (40%)	**34 (83%)**
	apple watch instead of Fitbit	2 (20%)	1 (2%)
	the task become more interesting	**6 (60%)**	**29 (71%)**
	shorter duration	**8 (80%)**	18 (44%)

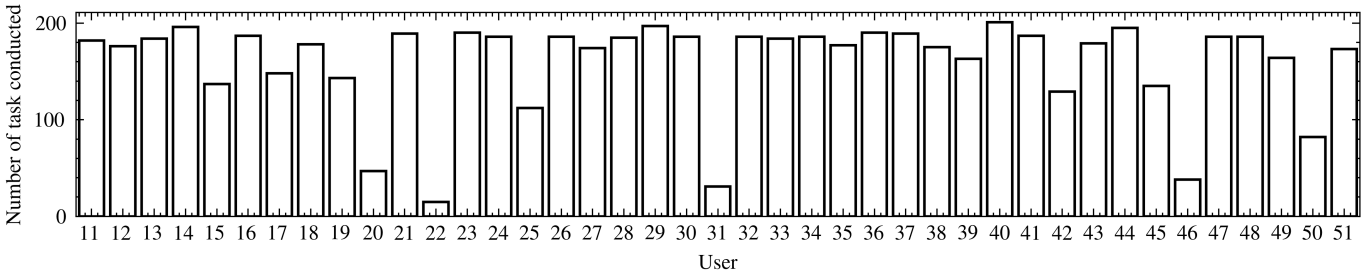


Fig. 8: Number of tasks conducted by each user in the user study.

- A limit of two valid completions per day was imposed for each task. After the limit is reached, task entry from the home screen will be disabled.
- The ranking screen was modified to display only the top 15 participants based on the number of task completions.

We conducted a user study involving 41 users to used the improved version of DailyExp in their daily life for one month. Both qualitative and quantitative analysis was performed.

A. Participants

In the study using an improved version of DailyExp, we recruited 41 participants (17 males and 24 females, aged 22-30) as app users, who are graduated students from the University of Tokyo. None of the participants from the preliminary study were permitted to take part in this subsequent study.

B. Procedure

The procedure of the orientation and instructions were almost the same as those conducted in the preliminary user study, despite that, we informed participants this time that there is a daily limit of two valid task completions for each type of task.

C. Quantitative Evaluation

1) Data Collected: We obtained a total of 6461 rounds of cognitive performance data, with 2115 rounds for the spatial continuous 2-back task, 2329 rounds for the Stroop task, and 2017 rounds for the FluidIQ task.

2) User Engagement: For the ratio of active users, as shown in Fig. 7, 38 out of the 41 users (other than user 22, 31, and 46) actively engaged with DailyExp. Consequently, the ratio of active users in the user study was 93%, showing a large improvement from that in the preliminary user study (60%).

Considering the number of tasks conducted, the average completion is approximately 157 tasks per user, which is three times the preliminary study's average of 55 tasks per user. For details, please refer to Fig. 8. We performed Mann–Whitney U test to compare the two groups of the number of tasks conducted. As results shown in Fig. 9, the group in the user study utilizing improved DailyExp showed a significant increase ($p=0.0008$) in the number of tasks conducted than that in the preliminary user study using the alpha version of DailyExp.

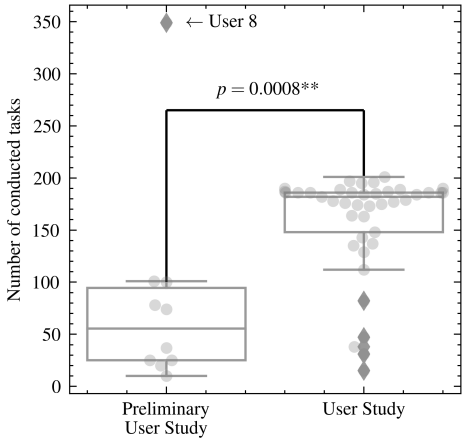


Fig. 9: A comparison of the number of conducted tasks between preliminary user study and user study using improved DailyExp

D. Qualitative Evaluation

1) Reason of losing motivation: In the user study, while the majority of the users sustained in task execution (85%), several users reported that they forgot to open the app and the experiment itself (12%).

2) Functionality that increases motivation: Similar to the result in the preliminary study, the Ranking screen was highly rated as contributing to increased motivation for task execution (78%). It is implied that a social competition atmosphere is a robust engaging feature and had a notable impact on the participants' motivation.

3) Possible improvements: The top suggested improvements were more monetary rewards (83%) and more interesting tasks (71%).

VI. Discussion

The results of our study offer several insights into enhancing user engagement and motivation in mobile-based cognitive tasks. Firstly, the significant improvement in user engagement, from 60% to 93% of active user ratio and three times more task completion per user, indicated that the two modifications made to the app were effective in enhancing user motivation.

On the one hand, the implementation of a limit of two valid completions per day for each task appears to align with Locke's goal-setting theory [10]. In Locke's research, the effect of goal setting on motivation was emphasized, and it was confirmed that setting clear goals, which the individual accepts, leads to better performance compared to ambiguous goals.

On the other hand, the ranking screen revealed to be the most effective feature in the preliminary study, was enhanced by concealing information about inactive users. This enhancement likely contributed to the increased motivation by creating a more competitive environment and providing a clearer sense of progress and achievement for active users. The mechanism behind this effect can be attributed to the psychological principle of social comparison, especially when their performance is visible and comparable. Thus, the concealment of inactive users may have heightened the perceived competition among active users, driving greater engagement and motivation.

VII. Limitations and Future work

It had to be addressed that we did not evaluate the two new features individually, meaning that the observed improvement in user engagement is likely an overall effect of both the introduction of achievable goals and the concealment of inactive users in the Ranking screen. This suggests that future studies should consider evaluating new features separately to better understand their impact on user motivation.

In future works, we plan to expand the coverage of cognitive aspects by administrating more cognitive batteries and implement a web-based dashboard for experimenters, which would allow them to easily adjust system factors and design their experiments. We expect DailyExp to be a useful tool for creating a large-scale real-world cognitive performance and physiology database. This database has great potential to contribute to the field of cognitive science by providing valuable information for understanding individual differences, intra-personal fluctuations, and the embodied nature of cognitive processes. Potential research questions include studying the impact of human rhythms on cognitive processes across different timescales (e.g., daily circadian rhythm, monthly menstrual cycle) and identifying biomarkers of cognitive processes correlated with physiological features.

VIII. Conclusion

In this paper, we presented DailyExp as a comprehensive tool for collecting cognitive performance and physiological data in everyday life settings. Building upon the alpha version published in [1], we implemented two key improvements. Firstly, we enhanced the sense of achievable goals by limiting users to two valid completions per day for each task. Secondly, we improved the Ranking screen by concealing information about inactive users.

We conducted a one-month user study with 41 individuals, the results indicated that these updates effectively enhanced user engagement. Our study demonstrated the app's effectiveness as a practical smartphone application for conveniently collecting data in daily life settings, showing consistent usage by a significant portion of users and successful data collection across multiple tasks.

Acknowledgment

This work was supported by JST SPRING, Grant Number JPMJSP2108.

References

[1] Xianyin Hu, Yuki Ban, and Shin'ichi Warisawa. "DailyExp: A Tool for Collecting Cognitive Performance and Physiological Data in Daily Life with Engaging Behavioral Design." The Fifteenth International Conference on Advanced Cognitive Technologies and Applications, June 2023.

[2] Wolfgang Tschacher, and Jean-Pierre Dauwalder, eds. ”Dynamical Systems Approach To Cognition, The: Concepts And Empirical Paradigms Based On Self-organization, Embodiment, And Coordination Dynamics”. Vol. 10. World Scientific, 2003.

[3] Susan Jongstra, Liselotte Willemijn Wijsman, Ricardo Cachucho, Marieke Peternella Hoevenaar-Blom, Simon Pieter Mooijaart, and Edo Richard. ”Cognitive testing in people at increased risk of dementia using a smartphone app: the iVitality proof-of-principle study.” JMIR mHealth and uHealth 5, no. 5 (2017): e6939.

[4] Zoë Tieges, Antaine Stíobhairt, Katie Scott, Klaudia Suchorab, Alexander Weir, Stuart Parks, Susan Shenkin, and Alasdair MacLullich. ”Development of a smartphone application for the objective detection of attentional deficits in delirium.” International psychogeriatrics 27, no. 8 (2015): 1251-1262.

[5] Tilman Dingler, Albrecht Schmidt, and Tonja Machulla. ”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, no. 3 (2017): 1-15.

[6] Pegah Hafiz, and Jakob Eyvind Bardram. ”The ubiquitous cognitive assessment tool for smartwatches: design, implementation, and evaluation study.” JMIR mHealth and uHealth 8, no. 6 (2020): e17506.

[7] Xianyin Hu, Shinji Nakatsuru, Yuki Ban, Rui Fukui, and Shin’ichi Warisawa. ”A physiology-based approach for estimation of mental fatigue levels with both high time resolution and high level of granularity.” Informatics in Medicine Unlocked 24 (2021): 100594.

[8] J. Ridley Stroop. ”Studies of interference in serial verbal reactions.” Journal of experimental psychology 18, no. 6 (1935): 643.

[9] Edith Lavy, and Marcel Van den Hout. ”Selective attention evidenced by pictorial and linguistic Stroop tasks.” Behavior Therapy 24, no. 4 (1993): 645-657.

[10] Edwin Locke, and Gary Latham. ”Goal-setting theory.” In Organizational Behavior 1, pp. 159-183. Routledge, 2015.