university of groningen
faculty of mathematics and natural sciences
artificial intelligence

Measuring Correlation of Cognitive Functions and Depression using App-based Game

Bachelor’s Project Thesis

Ievgen Teliatnykov, s3547272, i.teliatnykov@student.rug.nl

Supervisors: Dr. M.K. (Marieke) van Vugt

Abstract: Depression is one of the most common mental illnesses. Depression refers to a prolonged state of low mood and decreased motivation. Insomnia and fatigue are common physical symptoms of depression. Interestingly, depression is even visible in smartphone behaviours. Depressed people tend to send more text messages, spend more time on a smartphone, lower time spent outdoors, and receive shorter and fewer calls than non-depressed people. Moreover, depression negatively affects executive functioning and working memory. The research aims to find out if game performance correlates with mood fluctuations, individual differences in depression and smartphone behaviour. We expect that people with more depressed moods will have worse reaction time, concentration, working memory, and, thus, worse game performance. Depressed moods will be evaluated based on self-reports and smartphone behaviour. A 14-day study was conducted during which participants played a game that tested reaction time, concentration, and working memory. Participants also had to self-report their moods twice a day before playing the game to measure the mood fluctuations. To check the individual differences in depression, participants had to complete a questionnaire form at the study's start and end. Another app, called Behapp, tracked the user's behaviour (GPS location, number of phone calls, messages, and more) on the phone during the study. The results have shown that more calm people tend to remember more rules. However, it was also found that more calm, fulfilled, and concentrated people were completing levels at a slower rate than those who were less calm, fulfilled and concentrated. We did not find effects of individual differences in depression and smartphone behaviour on the performance variables. The potential reason for those results is a sample bias. Around 35% of participants have left during the data collection because the study may have been too overwhelming.

Key words: attention, concentration, depression, mood, smartphone behaviour, smartphone game, working memory.

1. Introduction

Depression is considered as one of the most common illnesses of mental health. In the United States, this illness is affecting more than 25 percent of adults. Depression is not only a mental disorder. It also affects the well-being and physical health of a human. Insomnia, weight fluctuations, increased pain sensitivity, fatigue, lower interest in sex, and weakened immune system are the common symptoms of depression (Goldman, 2019). Moreover, the research of Christopher and MacDonald (2005) has found that depression affects the allocation of attention and working memory. In the experiment, the participants with this mental disorder showed more impairments in phonological and visuospatial tasks than the control group without symptoms of the disease.

The problem with detecting depression is that people who experience that mental illness tend to be more socially closed and therefore are not visiting doctors about the issue (Elmer & Stadtfelt, 2020). A potential solution to this problem are smartphones. The study of Baqer (2017) has shown that there exists a negative correlation between happiness and depression. Therefore, happier people tend to have a lower probability of getting depression. It means that it is possible to find if a person is depressed by tracking their mood.

Smartphones do not require social interactions and have many mood-tracking applications, each with its unique analysis. Some of the applications are aimed to measure depression directly. (Caldeira et al., 2017).

Additionally, there are other methods of finding depression that can be done on a smartphone. One of the methods is through voice analysis. Depressed people tend to have a range of their pitch and volume being dropped. They start to speak flatter, lower, and softer. There is an application called "Cogito" based on a machine learning algorithm that performs voice analysis of a person and outputs the voice produced by a depressed or non-depressed person. However, the accuracy of the program is not perfect, as it is about 75% (Smith, 2017).

Another method is through self-reports or questionnaires. This is considered as the most common method. The classic mood report consists of thirty questions about a person's performance and feelings from the past week. Each question is a statement, and participants must agree with this statement on a scale from one to five, where one is completely agreed, and five is completely disagreed. Here is an example question, "you are facing a lack of concentration". There are also other variations of questionnaires made for measuring depression levels for the past day. One of them is the Patient Health Questionnaire-9 (PHQ-9). That questionnaire is measuring the level of depression. It was measured that the accuracy of PHQ-9 in identifying depression level is approximately 90% (Levis et al., 2019).

Moreover, depression can also be found by observation of a person's behaviour. The experiment of LiKam et al. (2013) has created an app, "MoodScope," that collects daily mood reports of participants and tracks their communication history. It was found that people who reported more depressive moods have sent more messages per day than those who had a less depressive mood. Additionally, it was also found that people with the symptoms of that illness tend to spend more time on their smartphones and receiving fewer and shorter calls (Razavi et al., 2020). Another observational experiment has shown that depression also affects time spent outdoors and sleeping hours. People with more symptoms of depression are oversleeping or undersleeping and spent fewer hours outside their home compared to people without symptoms (Doryab et al., 2014).

There is a smartphone game called "Wollie," which tests the cognitive performance (working memory, task-switching, and focusing skills) of a player. Wollie is an updated version of another game called "Rules" from the App Store (IOS). However, Wollie is only accessible in the Play Market (Android). The initial aim of the development of that application was to measure

how participants will learn while playing and check the effect of learning on other tasks (Doesburg & Taatgen, 2016).

Depression impairs executive functioning and working memory (DeBattista, 2005). Since Wollie requires working memory and executive functioning in order to progress, it is expected that people with depression are going to show poor performance compared to non-depressed people. Therefore, it may be possible to predict the presence of depression based on game statistics without the visit to a doctor.

The study of van der Zwan (2020) aimed to measure depression based on Wollie game performance. The short self-reported mood questions were used to find the fluctuations in mood. Those were the level of concentration, calmness, self-worth and cheerfulness. The questionnaires were used to find the individual differences in depression. It was found that better performance in the Wollie correlates with higher concentration and self-worth. However, it was also observed that higher depression scores in the questionnaire led to increased performance, which was not expected since depression negatively affects working memory.

This study aims to replicate the study of van der Zwan (2020) with additional depression measuring methods. There are three methods in total. The first one is mood questions for measuring fluctuation in mood. On top of van der Zwan's (2020) mood questions, the talkativeness and energy mood questions will be used. The second method is questionnaires for measuring the depressive symptoms. This method will remain the same as in van der Zwan's (2020) study. The last method is measuring depression through smartphone behaviour (number of calls, duration of calls, time spent in social apps, time spent at home, and more).

In order to track smartphone behaviour, the application "Behapp" will be used. A general idea of Behapp is that it tracks a person's location, communication history, and more. This app can be a tool to measure a depression level of a participant through behaviour.

1.1. Research Question:

This leads to the research question:

Does depression correlate with the performance in a smartphone game?

This research question can be divided into three sub-questions:

1. Does game performance correlate with self-reported mood fluctuations?

2. Does game performance correlate with individual differences in depression?

3. Does game performance correlate with smartphone behaviour?

It is hypothesized that depression will negatively impact the performance in the smartphone game "Wollie" because the game is testing working memory while working memory is negatively affected by depression.

2. Methods

2.1. Participants

In the study were 31 participants (17 Female), with a mean age of 23 years and range

18-34 years. The participants were recruited online using the information brochure that can be found in figure **A.12**. All participants were fluent English speakers. No specific education level was required for participation. Initially in the study were 48 participants, 17 of them have left.

2.2. Materials

Materials for the research were classified as depression measures and performance measures. Depression measure materials were used in order to estimate the fluctuations in mood, individual differences in depression and smartphone behaviour of a participant. Performance measure materials were present to find the participant's performance level.

2.2.1. Depression measure

2.2.1.1. Mood questions

Before each game session of Wollie, participants were asked to answer twelve questions on a scale from one to five. Where one was for strongly disagreeing, and five was for strongly agreeing. Those questions were necessary to track participant's fluctuations of mood each day over the study period. The first eight questions were obtained from a crowd-sourcing study that created a data collection of experience sampling questions (Krieke et al., 2016). Questions that were correlating with depression were chosen. Those questions were used in the study of van der Zwan (2020). The other four questions were created by analyzing other depressive symptoms that were not present in the original eight questions. The questions were designed so that a participant could answer them with real-time feelings by asking the question with the initial phrase 'At this moment I feel …'. The list of the questions can be seen below:

1a. At this moment, I feel calm.
1b. At this moment, I feel stressed.
2a. At this moment, I feel cheerful.
2b. At this moment, I feel down.
3a. At this moment, I am able to concentrate.
3b. At this moment, I am easily distracted.
4a. At this moment, I feel my life is worth living.
4b. At this moment, I feel I fall short.
5a. At this moment, I feel energized.
5b. At this moment, I feel tired.
6a. At this moment, I feel talkative.
6b. At this moment, I feel uncommunicative.

For each pair of questions, there existed a positive and negative variant of the question. The presence of negative and positive question versions was necessary to avoid acquiescence bias researched by Podsakoff et al. (2003). The bias was that some people tend to always give the highest possible answers without even giving the proper attention to the question.

2.2.1.2. Questionnaires

The questionnaires were used in order to find the general individual differences in depression. In the study were used in total three questionnaires. The reason for using three different questionnaires is to increase the accuracy of predicting depression levels. The first questionnaire was the Patient Health Questionnaire (PHQ-9). It was used to measure the current level of depression. PHQ-9 consists of nine questions, where each should be answered on a scale from zero to three. (Where 0 was for "not at all" and 3 was for "Nearly every day"). The total score range of 0-9 meant that participants have almost no depression. The score from 10 to 19 showed a minor depression or major depression level. The test with a score higher than 20 indicated severe major depression. Example question: little interest or pleasure in doing things.

The second questionnaire was the Perseverative Thinking Questionnaire (PTQ). That one was needed to predict how likely the participant was about to get depression. This test asked to answer 15 questions ranging from 0 to 4 (Where 0 was for "Never" and was for "Always"). The total score of a questionnaire ranged accordingly from 0 to 60. Example question: my thoughts take up all my attention.

The last was the Cognitive Failures Questionnaire (CFQ). This test was used to measure gaps in attention. CFQ was made of 25 questions, and just like in PTQ, answers to questions ranging from 0 to 4 (Where 0 was for "Never" and 4 was for "Always"). The final score was in the range of 0 to 100. Example question: do you find you forget appointments?

The list of all questions from questionnaires are shown in **figures A.1 - A.3.**

2.2.1.3. Behapp

Behapp was used in the study to measure the participant's smartphone behaviour. Behapp was an application created by Jagesar (2021), that when installed on the smartphone, tracked the behaviour and communicative history of a participant. The app was able to track a person's location, time spent in each app, count the number of incoming and outgoing calls, and much more. Behapp was automatically formatting raw data and distributed them in three sections. Those sections were location, call, and app foreground usage. In total, Behapp provided 60 unique variables. For the current research, 6 variables from the total number of variables were chosen. The variables were chosen based on the connection to depressive behaviour. As more depressed people tended to have fewer and shorter calls, spend more time on a smartphone, spend more time at home and send more messages per day

From the location section:

1. Percentage of time spent at home.

As shown by Doryab et al. (2014), people with depression tended to spend more time at home.

From the call section:

1. Number of all calls.
2. Duration of all calls in seconds.

The duration and number of call variables showed how long and how often in general participants use calls. Razavi et al. (2020) showed that receiving shorter and fewer calls were connected with higher depression.

3. Number of missed calls.

The avoidance of the calls was also a sign of depressive symptoms.

From app foreground section:

1. Duration opened communication apps in seconds.
2. Duration opened social media apps in seconds.

LiKam et al. (2013) have reported that the more messages sent by a participant per unit of time, the more depressive a participant was. Since Behapp cannot track what exactly a participant was doing in the app (it was impossible to track the number of messages sent), the variables of time spent in the app were the closest to the number of messages sent per day.

2.2.2. Performance measure

2.2.2.1. Wollie

Wollie was an application used to measure the performance of a participant. The game measured executive functioning and working memory since it required participants to remember different rules. The goal of the game was to pass through as many levels as possible. Wollie has in total 18 levels. The first nine levels were in beginner mode. The other nine levels were in expert mode. Each level was made out of 10 sub-levels, wherein each sub-level, a new rule was introduced. To complete a level, participants must have remembered the rules given in each sub-level and tap the tiles in the order given by the rules. In the first sub-level, a first rule was given. The second sub-level already has a rule from the 1st sub-level and introduces a new rule from the current sub-level, and so on (the third sub-level has two rules from the first two sub-levels and the third one from the current one). Sub-level represented a four by four grid with 16 tiles in total. Each tile was featured by a number from one to ten with a random image. The image could be an animal, a monster, a car, and more. Additionally, each

image was represented by a particular colour. Here is an example of two rules from the game.

Rule 1: tap numbers in descending order

Rule 2: tap all things green.

If a participant is currently starting a level 1 game, the game shows a rule of 1st sub-level and a continue button (a rule-introducing screen). When a participant remembers a rule and presses the continue button, the first sub-level starts, and a 30-second timer plays. The participant starts pressing the tiles. If a tile was pressed correctly by the rule, the tile disappears. If a guess was wrong, the error sound is played, and the tile is flashed with red colour. There is a 0.5-second delay before another tile can be pressed, not to allow participants to press several tiles at once and guess them. When all tiles from the 1st sub-level have disappeared, a participant gets 14 more seconds to the timer, and the second rule-introducing screen appears, which shows a second rule. When participants have learned a second rule and pressed a continue button, the second sub-level starts. Now, participants must tap all the tiles according to the second rule. When all the tiles representing a second rule have disappeared, the game shows a small text on the top of a screen that asks the player to start tapping the tiles according to the first rule (the game does not show a rule directly, it shows a number of a rule), which was initially showed before the play of first sub-level. If a player reaches a tenth sub-level, the player must already remember ten rules in order to complete it. When the last sub-level of the first level is completed, the second level is unlocked. The second level can now be accessed from the level-choosing screen. If the timer reaches zero seconds, then the game stops, and the player loses a level. Each new level introduces a harder set of rules which can be encountered. When the nine levels of beginner mode are finished, the expert mode is unlocked. Expert mode has the same rules but has less time to finish it. Expert mode gives only 20 seconds to finish a level, with only an addition of nine seconds per sub-level completed. The screenshots from the Wollie game can be seen in **figures A.4-A.7**

The Wollie game has multiple performance measures. Four of them were chosen.

1. Maximum level reached

The maximum level reached was a variable that represented the level that participants have reached during the session. It indicated how fast participants were progressing through the game.

2. Accuracy of correctly pressed tiles

Accuracy showed the percentage of correct moves participants made during the session. The correct move was represented each time when the participant correctly guessed a tile-based on the given rule. An incorrect move was when the wrong tile was pressed.

3. Number of rules remembered

The rules remembered variable represented a number of rules a participant has remembered during the session.

4. Percentage of completed rules

The percentage of completed rules was measured by dividing rules remembered by their sum with forgotten rules. The rule was considered forgotten if participants guessed a wrong tile three times in a row. There were several reasons for guessing the wrong tile, such as not finishing looking among all the tiles or just a simple misclick. Therefore participants have two tries before the rule was considered forgotten

2.3. Procedure

2.3.1. Participant recruitment

The information brochure was published online that describes requirements, tasks, and rewards for participation. The contact details of one of the researchers were present in a brochure.

The requirements were:

- Android smartphone with at least Android version 5.0 (Lollipop).
- Age should be between 18-35
- Not colour blind
- Not working late night shifts
- Available for 14 days in a row without expecting large variations in the daily routine.

Tasks were:

- Install two apps on the smartphone.
- The first app was tracking the behaviour (the time spent outside and inside the home, number and duration of calls, time spent on the communication and social apps - all info was superficial, so it is not possible to look into what exactly participant was doing in those apps).
- The second app was a game.
- A participant must have played the game for 5 - 10 minutes twice a day for a period of 14 days and fill a mood report before each session.
- Participants were asked to fill in questionnaires at the start and the end of the research.

Rewards were:

- A 20 Euro money compensation.

- Personalized report about participant's behaviour and performance.

. When interested subjects contacted the researcher, they received and filled an online informed consent form where they agreed with the requirements and the tasks. Participants were not informed about the topic and research question of the research.

2.3.2. Pre-test

Recruited participants were asked to fill in three questionnaires. When questionnaires were completed, participants provided researchers with their Google accounts in order to give them access to the Wollie application, as access to the application was private. When access was given, participants received a video explaining how to set up Behapp and Wollie. Video also asked participants to set two alarms that reminded players to play Wollie. In both apps, participants were asked to give permission to external storage in order to track the data. That

data was later sent to the researchers and Behapp. By the end of the pre-test, participants had to read through game instructions in order to learn how to play.

2.3.3 Test

On the first day of the experiment, participants had time to finish the installation of apps and practice 5-10 minutes before the next day starts with the research. When the actual test started, participants were required to play Wollie two sessions per day for 5-10 minutes each. If a participant has lost the game by playing it only for 2 minutes, the participant must have played Wollie again until the player reached a minimum time of play of 5 minutes. After that, the participant may have exited the game until the next session or continue playing until the maximum time of 10 minutes has been reached. Participants knew how much time they must play by the progress bar on the main screen. When a player left the game, progress was saved on the smartphone. The break between the two sessions must have been at least two hours. Wollie has not allowed to play a game if the first session of the game started at 23:00, as the next session would already occur on the next day. Therefore, it was recommended to play Wollie in the morning and afternoon. The sessions which were not finished were not recorded in the research data. Each night at 5:00, Wollie was attempting to store the game log files on Google Cloud Storage of Firebase, where the results of each participant can be seen by the researchers. If it was detected that one of the participants was not active, that was a sign that the participant had left the playing.

2.3.4 Post-test

The participants who played Wollie for all 14 days have received questionnaires which they have completed in the pre-rest to complete again. When questionnaires were completed for the second time, participants provided the bank details and received the 20 euro compensation. Additionally, they have received the personalized report with their own performance and behaviour.

2.4. Data analysis

Data analysis will be done with the help of RStudio, based on the programming language R.

Data of participants who have not finished testing and left during the experiment was removed.

In order to find the effects of depression on the performance in Wollie, the ANOVA statistical test was chosen. Three separate ANOVA models were used to analyze the effect of mood fluctuations, individual differences in mood and smartphone behaviour on executive functioning and working memory using the game performance data.

2.5 Pre-processing

2.5.1 Mood questions

In order to analyze the fluctuation in mood, it is necessary to look into responses to mood questions.

The questions had a positive and negative version of themselves. Therefore, the Spearman correlation test was done to ensure that participants were answering questions honestly. The correlation test can be seen in **figure A.8**. The figure shows that negative and positive versions of questions are significantly correlating. Since correlation is present, both versions of questions were combined to simplify the statistical analysis. The combination was done by reversing the negative version's self-reported scores and averaging it with a positive version of a mood question.

Another spearman correlation test was done, where mood questions were compared with each other to avoid a possible multicollinearity problem in the data analysis. The test is shown in

figure A.9. As it can be seen from the correlation test, questions do not correlate with each other.

2.5.2 Questionnaires

The correlation test was done to check if pre-test and post-test questionnaires data are correlating.

For a PTQ test, [t(31) = 3.07, p = 0.004].

For a PHQ test, [t(31) = 6.11, p < 0.001].

For a CFQ test, [t(31) = 8.04, p < 0.001].

Since the pre-test and post-test of each questionnaire have a strong correlation, the pre-data and post data were combined in one by averaging them.

The combined questionnaire data was tested in correlation with itself for the same reason of multicollinearity avoidance. The test can be seen in **figure A.10.** PTQ, PHQ and CFQ questionnaires were not correlating with each other.

2.5.3 Behapp

In the case of smartphone behaviour, the Pearson correlation test was chosen because Behapp has interval variables, while mood questions and questionnaire scores were made of ordinal variables. The results of the test can be seen in **figure A.11**. It is shown that the number of calls correlates significantly with the duration of all calls, with the number of missed calls and the percentage of time spent at home. Additionally, the duration of all calls is significantly correlating with the number of missed calls. Therefore, in order to avoid the multicollinearity in the smartphone behaviour ANOVA model. The number of calls and number of missed calls were removed from the model.

3. Results

3.1 Fluctuations in mood

ANOVA test was done to find whether there exists an effect of mood question score on the Wollie performance. The results of the test are shown in **figure B.1**. The test showed that there is a variable with significant effect on the number of remembered rules, that is a calmness question score [t(1, 822) = 6.45, p = 0.011], and three variables that have an effect on the maximum level reached per session, those are calmness [t(1, 822) = 9.25, p = 0.0025], concentration [t(1, 822) = 4.84, p = 0.0278] and fulfillness [t(1, 822) = 6.31, p = 0.0122]. The maximum level reached during the whole experiment has not shown any significant variables, as well as accuracy of correctly pressed tiles and percentage of completed rules.

The effect of the calmness question score on the number of remembered rules is shown in **figure 1**.

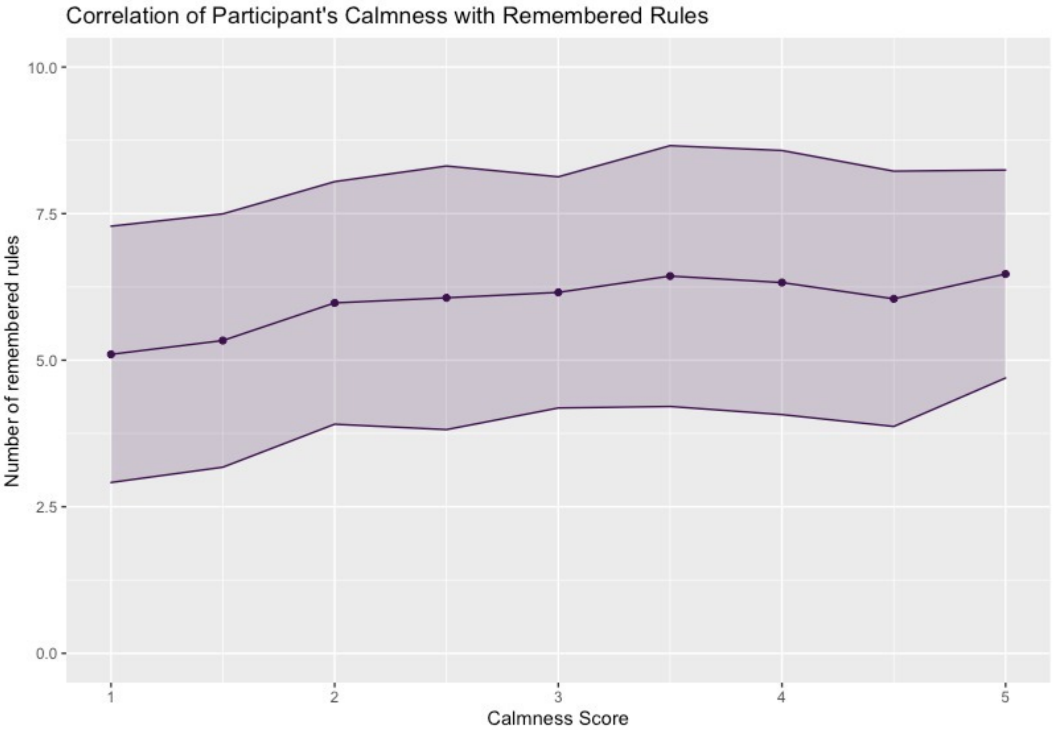


Figure 1: A graph with a mean number of remembered rules over the calmness mood question score

The graph shows that people who picked a higher score for the calmness question have remembered, on average, more rules per session compared to ones who picked less. Those results line up with the hypothesis, as it was expected to have better performance when people feel calm and not stressed.

The effect of calmness, concentration, and fulfillness questions on the maximum reached can be seen in **Figure 2**.

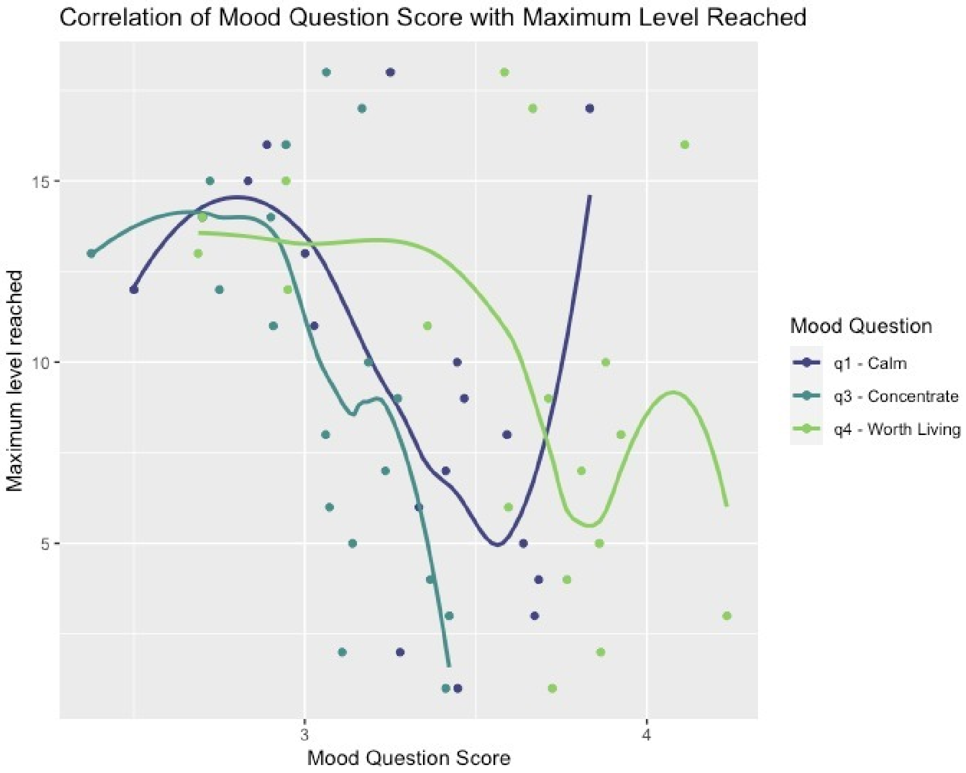


Figure 2: A graph that shows the mean maximum level reached per scores of calmness, concentration and fulfillness questions.

The graph shows that participants who feel more fulfilled and more concentrated are completing levels slower than those who reported feeling not worth living and being distracted. The same results are for the calmness score. The less calm the participant is, the slower participant is completing the game. However, the calmness line on the graph contains a spike in the end. This is because only one of the participants was able to score a higher level on a higher mood question score. Those results do not fit the hypothesis, as it was expected that people who are calmer, concentrated, and more fulfilled would perform better with higher mood scores. We got opposite results. More calm, concentrated and fulfilled participants were performing worse with higher mood scores.

Moreover, the contradiction was obtained. It was shown that higher calmness scores increase the number of remembered rules and decrease the speed of completing the levels.

3.2 Individual differences in depression

There were, in total, three questionnaires that participants did to measure the individual differences in depression. Those are PHQ, PTQ and CFQ.

The ANOVA test was made to rate the significance of the questionnaire's effect on the performance variable. The results of the test can be seen in **figure B.2**. As can be seen, none of the questionnaire scores had a significant effect on any performance variable. Since no significant impact was found, questionnaire scores have not shown that current depression level (PTQ), the likelihood of getting depression (PHQ) and gaps in attention (CFQ) negatively affects the performance in a game.

3.3 Smartphone behaviour

Behapp is a behaviour tracking app that was installed on the smartphones of the participants to measure the participant's smartphone behaviour. The ANOVA test was made to determine if the significant effect of Behapp variables was present on the performance measure variables. The impact of Behapp variables on performance variables is shown in

figures B.3-B.4. It was found that there were no significant effects of Behapp on Wollie performance. Therefore, it is shown that the number of calls, missed calls, the number of text messages sent, time spent on the phone, and time spent at home do not correlate with Wollie gameplay.

4 Discussion

4.1 Findings

This study aimed to determine whether game performance correlates with depressive symptoms. It was hypothesised that depression is associated with impairments in playing the game. That research question was divided into three sub-questions. The first sub-question is about how game performance correlates with mood fluctuations. We assumed that people with more calm, cheerful, concentrated, fulfilled, energised and talkative (overall less depressed) behaviour will show better game performance. The obtained results were both rejecting and confirming the hypothesis. The participants who reported themselves with more calm behaviour have remembered more rules per session. However, participants with the same score for calmness have progressed through the game slower (fewer levels completed per session) than those who were less calm. Those results contradict each other since participants who remembered more rules should have progressed through the game faster, as it is needed to remember ten rules to progress to the next level. It turns out that those participants who have remembered more rules, on average, were most of the time away from finishing the level. This can be explained by the fact that each level is more complex than the previous one. Those who initially progressed through the game faster than the others have met the rules that are harder to remember. Therefore, they started to remember fewer rules on average than those who are still on an easier level.

Moreover, players who considered themselves more concentrated and fulfilled have progressed through the game slower than those who were less concentrated and fulfilled. This is interesting, especially with concentration, since concentrated participants should perform better than non-concentrated. Our findings instead showed that participants who felt more concentrated performed worse than participants who reported to be less concentrated. The study made by Zwan (2020), which we are replicating, has received different results. The participants in the original study have remembered more rules, completed more rules and had higher accuracy of correctly pressed tiles with higher concentration. While in this study, no effect of concentration was found on those variables.

The second sub-question was about the correlation of game performance with individual differences in depression. The individual differences in depression were measured with three questionnaires, which are PTQ, PHQ and CFQ. It was expected that people who had higher scores in those questionnaires would have shown worse results in the Wollie. However, the questionnaires did not have any effect on the performance variables. In terms of PHQ, in our study, we had three participants with a severe level of depression, nine moderately severe, eleven moderates, four mild and four minimal. Generally, the study had participants with all levels of depression severity. One of the possible reasons for not finding the effect is the fact that we obtained data from a limited number of participants (31). The original study of this experiment by van der Zwan (2020) has got a significant effect. However, the effect was the opposite of what was expected. People who had higher scores of PHQ have remembered, completed more rules and reached a higher level of the game. One of the possible explanations for the different results is the percentage of depressive people in the experiment. The van der Zwan's (2020) experiment had only five severely or moderately severe participants out of 49 (10%). While in this study were nine severely

depressed participants out of 31 (29%). The increase in the depression level of this study can be explained by the study of Bueno-Notivol (2021), which states that depression prevalence during the COVID-19 pandemic has increased by seven times. The original study was conducted when the COVID-19 just started. Therefore, depression prevalence was not as high as in 2021. However, at the same time, it could also be a sampling bias

The last sub-question of this study is about game performance correlation with smartphone behaviour. It was expected that people who stay more at home, have fewer and shorter mobile calls, send fewer text messages and spend more time on their phones are more depressed. Therefore, those people should have shown worse performance in the game than those who have opposite behaviour. The Behapp application was used to track the smartphone behaviour of participants (time spent at home, time spent in social media/communication apps, number of missed calls, number and duration of calls). The results of this study have not shown any significant effect of phone-derived social contact variables on the game performance. The reason for not finding the effect is that Behapp was not supporting all Android manufacturers. Therefore only twenty participants out of thirty-one were able to set Behapp working correctly. Not only the lack of participants was the reason. Behapp requires an internet connection in order to keep track of the data. We asked our participants to always be connected to the internet. However, we observed that many participants were missing internet connection, which has led to a loss of smartphone behaviour data.

4.2 Limitations

One of the limitations of this study is that it is too overwhelming. We asked our participants to answer 24 questions and play the game for 10-20 minutes per day in the period of two weeks. Additionally, they were always required to be connected to the internet. Due to those overwhelming requirements, participants may have begun to not give a serious answer to the mood question or stopped to pay attention to the game itself. That also led to the session skipping. There were 109 skipped sessions in total out of 744. On top of that, 17 out of 48 participants (35.4 %) dropped out of the experiment.

Another limitation of the study is the different weight of the performance measures. In the game, there are a total of 18 levels. Half in the beginner mode and half in the expert mode. The game measures the percentage of correctly pressed tiles, the number of remembered rules, percentage of remembered rules. The problem is, for example, 80% of correctly pressed tiles in level 1 and 80% of correctly pressed tiles in level 18 are entirely different results. Since level 18 has more complex rules than level 1 and, therefore, requires more attention and working memory. However, in the experiment, the game performance from different levels was considered on the same scale. So 80% of accuracy in levels 1 and 18 had no difference in the experiment's results analysis. The consideration of each level's difficulty would have made the analysis of the data complicated.

4.3 Future work

The possible improvement for the experiment is to create one universal level for the game to avoid the different weight of the performance measures. The player can play the game until the player loses it. For example, a game has a pool of 30 rules, and each rule is chosen randomly.

The four by four grid of tiles will not be enough for 30 rules. Therefore, a five by five grid will be created. Additionally, the player will only be allowed to make a mistake three times, or the game will end. This will not allow players to pick tiles randomly and will eliminate the luck in the game. In this case, it will be hard to reach the end of the game since remembering 30 rules is almost impossible. Players are always going to start from the same level of difficulty. Therefore the weight of the performance variables will be the same. Moreover, game performance variables such as the number of remembered rules, percentage of completed rules and accuracy of correctly pressed tiles will be unnecessary. Since now, the game performance can be only judged by the last level reached. That is going to simplify the analysis of the result. This possible improvement to the game could show more results that are lining up with the hypothesis.

References

Baqer, G. (2017). Correlation between depression and happiness among Kuwait university students. *European Psychiatry*, *41*(S1), S522. https://doi.org/10.1016/j.eurpsy.2017.01.692

Bueno-Notivol, J. (2021). Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, *21*(1). https://doi.org/10.1016/j.ijchp.2020.07.007

Caldeira, C., Chen, Y., Chan, L., Pham, V., Zheng, K. (2018). Mobile apps for mood tracking: an analysis of features and user reviews. *Annual Symposium proceedings. AMIA Symposium*. 495-504.

Christopher, G., & MacDonald, J. (2005). The impact of clinical depression on working memory. *Cognitive neuropsychiatry*, *10*(5), 379–399. https://doi.org/10.1080/13546800444000128

DeBattista C. (2005). Executive dysfunction in major depressive disorder. *Expert review of neurotherapeutics*, *5*(1), 79–83. https://doi.org/10.1586/14737175.5.1.79

Doesburg, I., & Taatgen, N. (2016.). Using a smartphone game to promote transfer of skills in a real world environment.

Doryab, A., Min, J. K., Wiese, J., Zimmerman, J., & Hong, J. I. (2014). Detection of behavior change in people with depression.

Elmer, T., Stadtfeld, C. (2020).Depressive symptoms are associated with social isolation in face-to-face interaction networks. *Sci Rep* 10, 1444. https://doi.org/10.1038/s41598-020-58297-9

Goldman, L. (2019, November 22). *What is depression and what can I do about it?* MedicalNewsToday. https://www.medicalnewstoday.com/articles/8933

Jagesar, R. (2021). *Behapp* [Mobile app for tracking behaviour]. Google play. https://behapp.org/

Krieke, L. V. D., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B., Emerencia, A. C., Schenk, H. M., Vos, S. D., Snippe, E., Wichers, M., Wigman, J. T., Et al. (2016). Hownutsarethedutch (hoegekisnl): A crowdsourcing study of mental symptoms and strengths. International journal of methods in psychiatric research, 25(2), 123–144.

Levis Brooke, Benedetti Andrea, Thombs Brett D. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis *BMJ* 2019; 365 :l1476

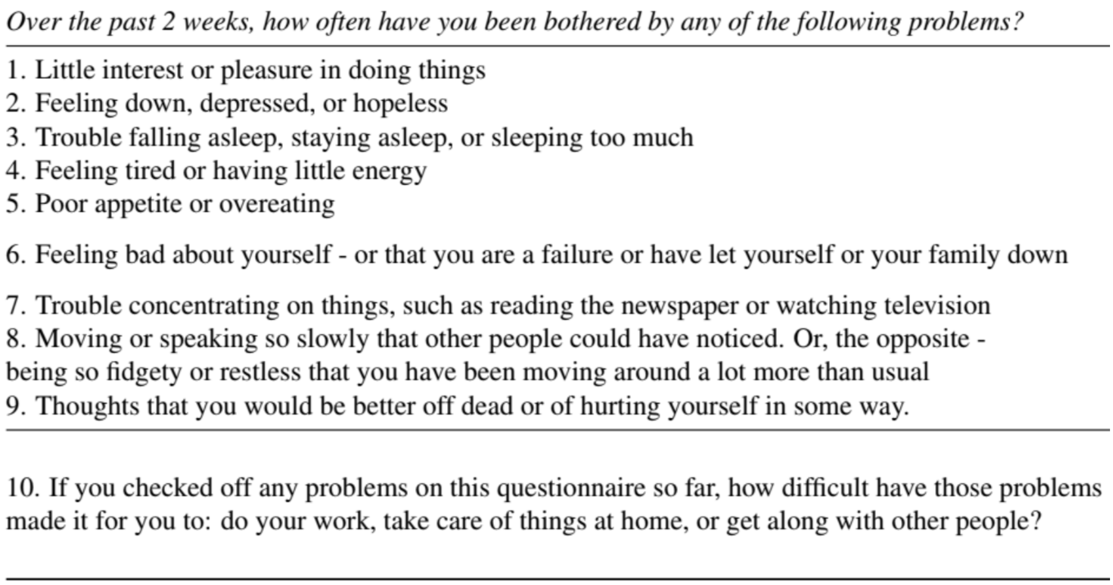
LiKamWa, R., Liu, Y., Lane, N. D., & Zhong, L. (2013). Moodscope: Building a mood sensor from smartphone usage patterns, In Proceeding of the 11th annual international conference on mobile systems, applications, and services.

Smith, D. (2017, October 12). Capturing the Sound of Depression in the Human Voice. KQED. https://www.kqed.org/futureofyou/435986/capturing-the-sound-of-depression-in-the-human-voic%20e

Razavi, R., Gharipour, A., & Gharipour, M. (2020). Depression screening using mobile phone usage metadata: A machine learning approach. *Journal of the American Medical Informatics Association*, 27(4), 522–530.

Zwan, P. T. (2020). *Tracking Cognition over Time with a Smartphone Game*. Artificial Intelligence - University of Groningen.

A.Appendices

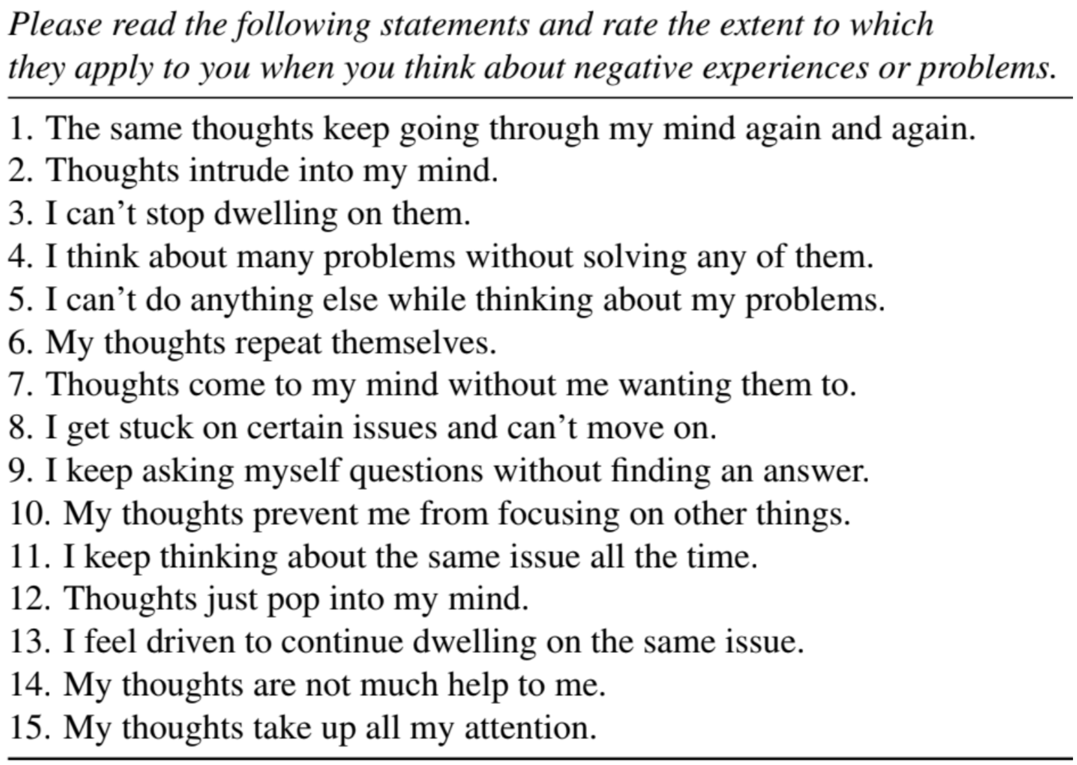


Over the past 2 weeks, how often have you been bothered by any of the following problems?

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling asleep, staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed. Or, the opposite - being so fidgety or restless that you have been moving around a lot more than usual
9. Thoughts that you would be better off dead or of hurting yourself in some way.

10. If you checked off any problems on this questionnaire so far, how difficult have those problems made it for you to: do your work, take care of things at home, or get along with other people?

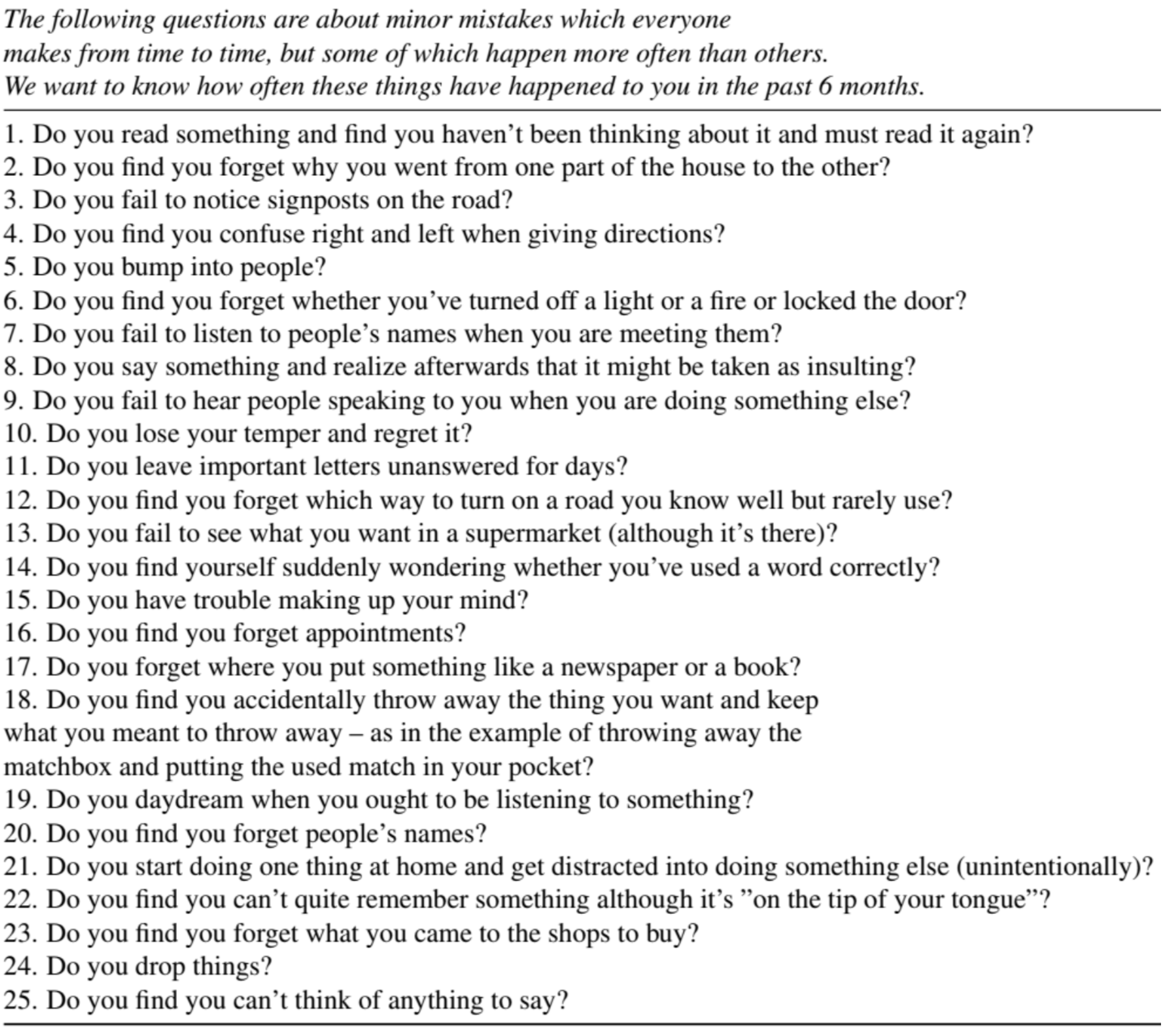
Figure A.1: A list of PTQ-9 questions



Please read the following statements and rate the extent to which they apply to you when you think about negative experiences or problems.

1. The same thoughts keep going through my mind again and again.
2. Thoughts intrude into my mind.
3. I can't stop dwelling on them.
4. I think about many problems without solving any of them.
5. I can't do anything else while thinking about my problems.
6. My thoughts repeat themselves.
7. Thoughts come to my mind without me wanting them to.
8. I get stuck on certain issues and can't move on.
9. I keep asking myself questions without finding an answer.
10. My thoughts prevent me from focusing on other things.
11. I keep thinking about the same issue all the time.
12. Thoughts just pop into my mind.
13. I feel driven to continue dwelling on the same issue.
14. My thoughts are not much help to me.
15. My thoughts take up all my attention.

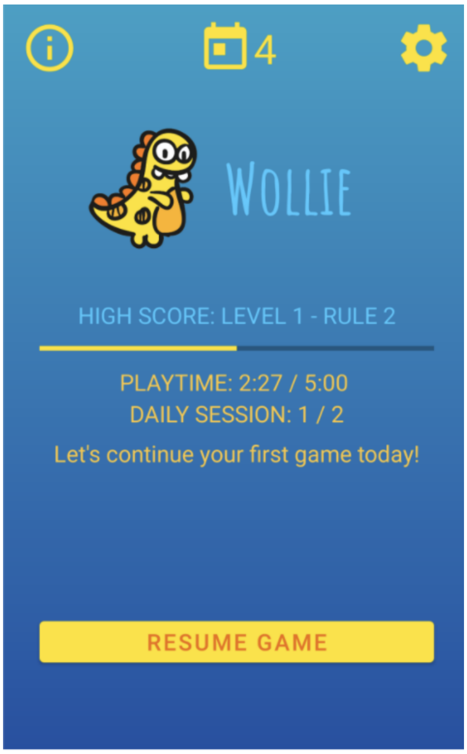
Figure A.2: A list of PHQ questions.



The following questions are about minor mistakes which everyone makes from time to time, but some of which happen more often than others. We want to know how often these things have happened to you in the past 6 months.

1. Do you read something and find you haven't been thinking about it and must read it again?
2. Do you find you forget why you went from one part of the house to the other?
3. Do you fail to notice signposts on the road?
4. Do you find you confuse right and left when giving directions?
5. Do you bump into people?
6. Do you find you forget whether you've turned off a light or a fire or locked the door?
7. Do you fail to listen to people's names when you are meeting them?
8. Do you say something and realize afterwards that it might be taken as insulting?
9. Do you fail to hear people speaking to you when you are doing something else?
10. Do you lose your temper and regret it?
11. Do you leave important letters unanswered for days?
12. Do you find you forget which way to turn on a road you know well but rarely use?
13. Do you fail to see what you want in a supermarket (although it's there)?
14. Do you find yourself suddenly wondering whether you've used a word correctly?
15. Do you have trouble making up your mind?
16. Do you find you forget appointments?
17. Do you forget where you put something like a newspaper or a book?
18. Do you find you accidentally throw away the thing you want and keep what you meant to throw away – as in the example of throwing away the matchbox and putting the used match in your pocket?
19. Do you daydream when you ought to be listening to something?
20. Do you find you forget people's names?
21. Do you start doing one thing at home and get distracted into doing something else (unintentionally)?
22. Do you find you can't quite remember something although it's "on the tip of your tongue"?
23. Do you find you forget what you came to the shops to buy?
24. Do you drop things?
25. Do you find you can't think of anything to say?

Figure A.3: A list of CFQ questions.



HIGH SCORE: LEVEL 1 - RULE 2

PLAYTIME: 2:27 / 5:00
DAILY SESSION: 1 / 2

Let's continue your first game today!

Figure A.4: Main menu and the progress bar.

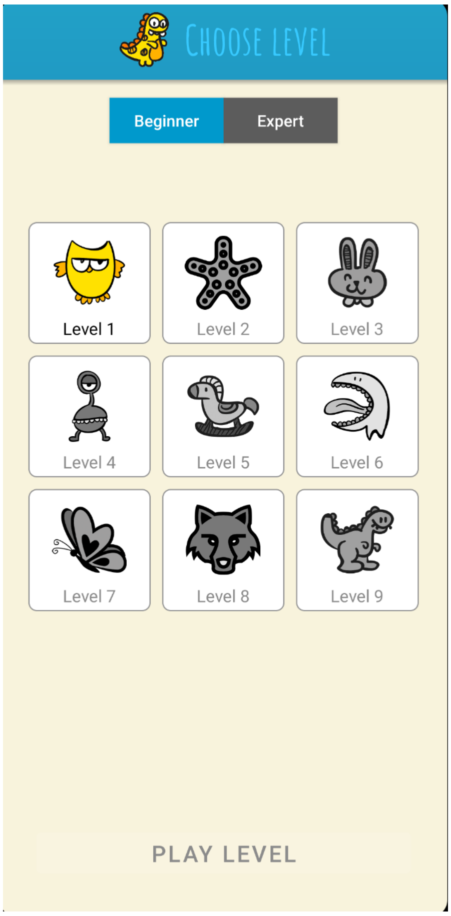
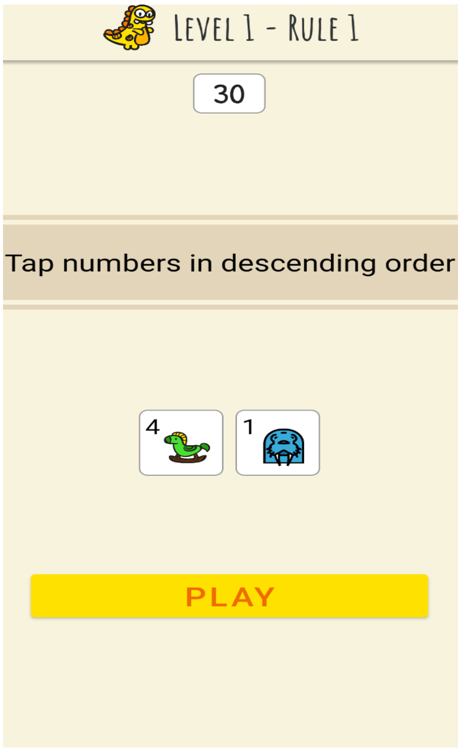
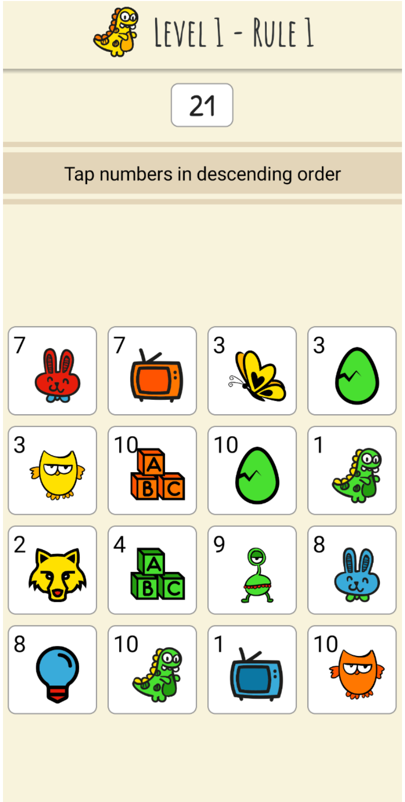


Figure A.5: The level choosing screen



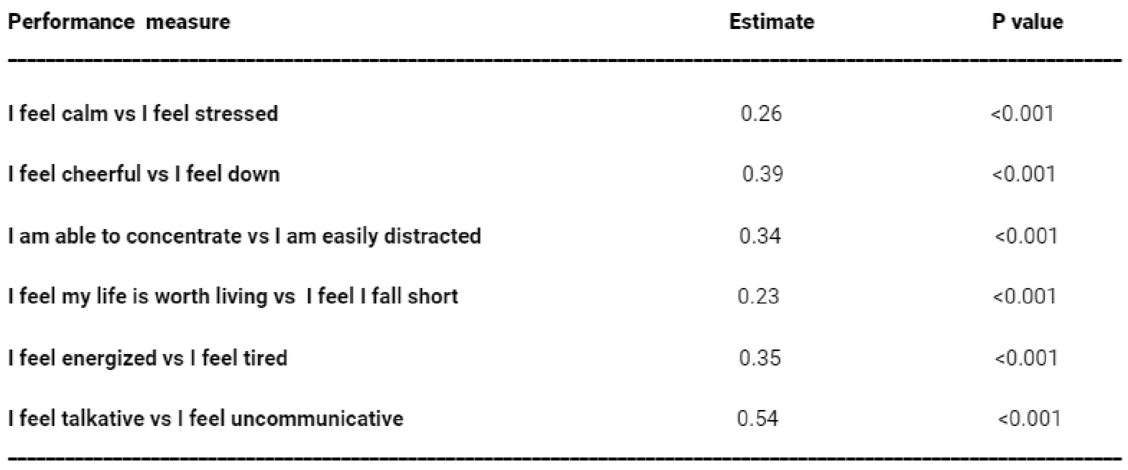
Tap numbers in descending order

Figure A.6: An introduction of the new rule screen



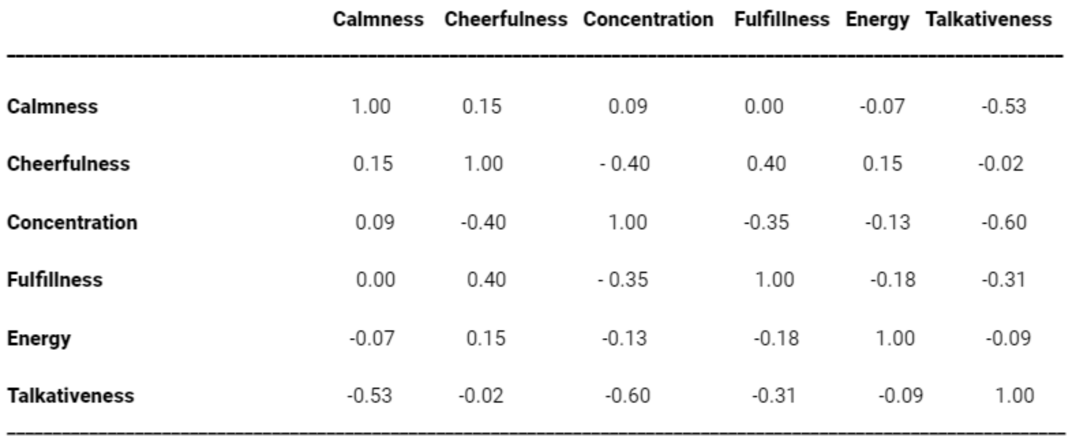
Tap numbers in descending order

Figure A.7: The gameplay screen



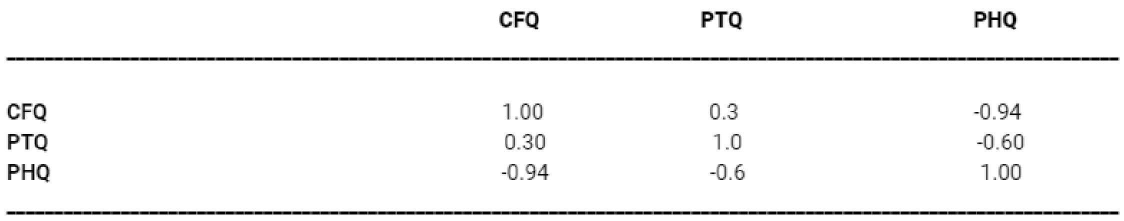
Performance measure	Estimate	P value
I feel calm vs I feel stressed	0.26	<0.001
I feel cheerful vs I feel down	0.39	<0.001
I am able to concentrate vs I am easily distracted	0.34	<0.001
I feel my life is worth living vs I feel I fall short	0.23	<0.001
I feel energized vs I feel tired	0.35	<0.001
I feel talkative vs I feel uncommunicative	0.54	<0.001

Figure A.8: A Spearman correlation test between negative and positive versions of mood questions.



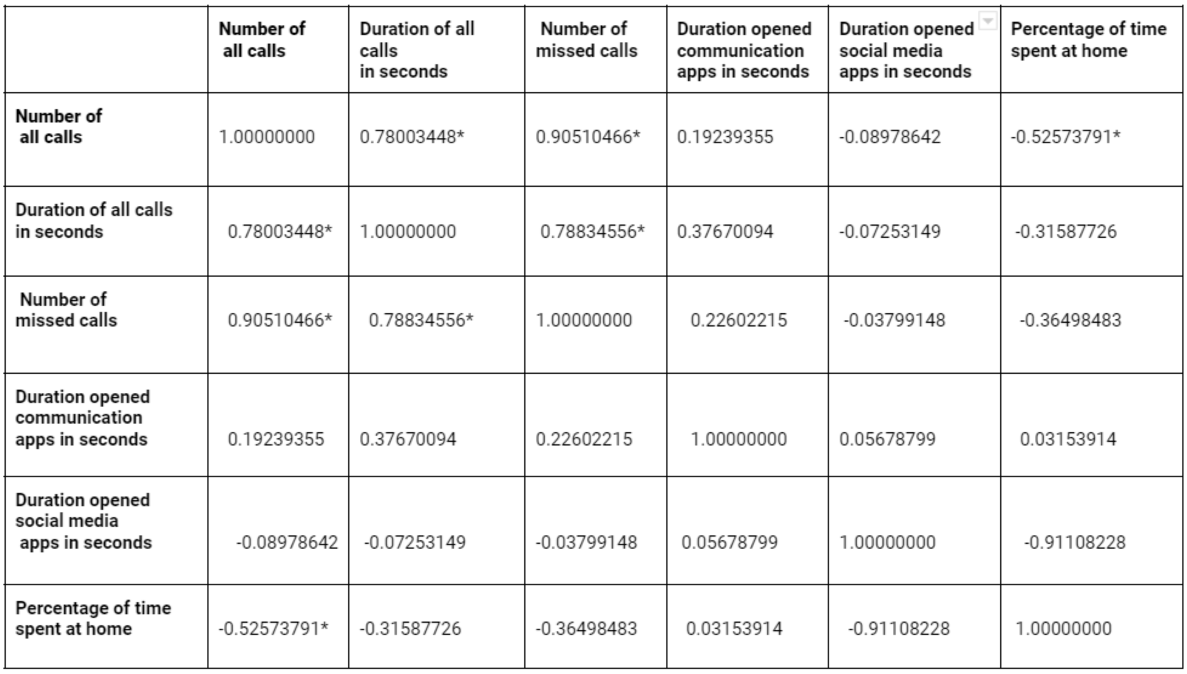
	Calmness	Cheerfulness	Concentration	Fulfillness	Energy	Talkativeness
Calmness	1.00	0.15	0.09	0.00	-0.07	-0.53
Cheerfulness	0.15	1.00	- 0.40	0.40	0.15	-0.02
Concentration	0.09	-0.40	1.00	-0.35	-0.13	-0.60
Fulfillness	0.00	0.40	- 0.35	1.00	-0.18	-0.31
Energy	-0.07	0.15	-0.13	-0.18	1.00	-0.09
Talkativeness	-0.53	-0.02	-0.60	-0.31	-0.09	1.00

Figure A.9: A Spearman correlation test of mood questions. *Shows the significance.



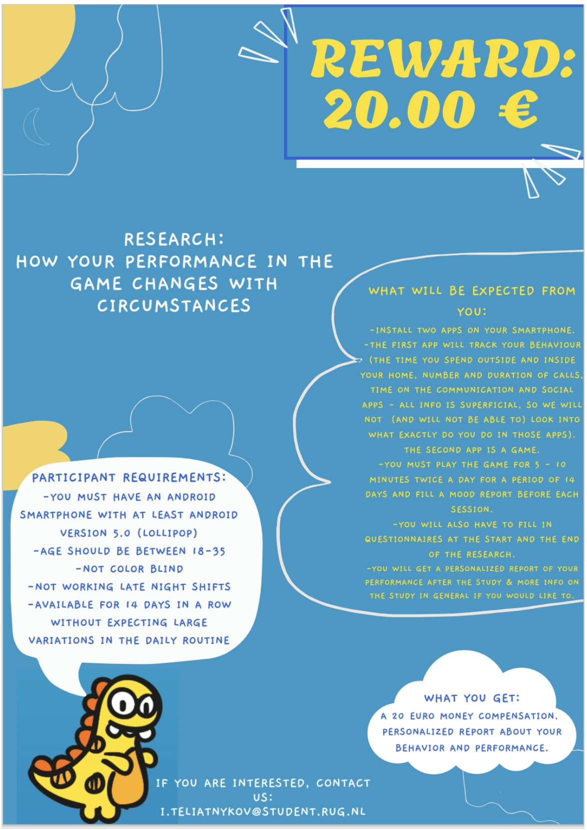
	CFQ	PTQ	PHQ
CFQ	1.00	0.3	-0.94
PTQ	0.30	1.0	-0.60
PHQ	-0.94	-0.6	1.00

Figure A.10: Correlation test of Spearmen, where questionnaires were compared. * Shows the significance



	Number of all calls	Duration of all calls in seconds	Number of missed calls	Duration opened communication apps in seconds	Duration opened social media apps in seconds	Percentage of time spent at home
Number of all calls	1.00000000	0.78003448*	0.90510466*	0.19239355	-0.08978642	-0.52573791*
Duration of all calls in seconds	0.78003448*	1.00000000	0.78834556*	0.37670094	-0.07253149	-0.31587726
Number of missed calls	0.90510466*	0.78834556*	1.00000000	0.22602215	-0.03799148	-0.36498483
Duration opened communication apps in seconds	0.19239355	0.37670094	0.22602215	1.00000000	0.05678799	0.03153914
Duration opened social media apps in seconds	-0.08978642	-0.07253149	-0.03799148	0.05678799	1.00000000	-0.91108228
Percentage of time spent at home	-0.52573791*	-0.31587726	-0.36498483	0.03153914	-0.91108228	1.00000000

Figure A.11: a Pearson correlation test. The Behapp variables were compared. * Shows the significance.



RESEARCH:
HOW YOUR PERFORMANCE IN THE GAME CHANGES WITH CIRCUMSTANCES

WHAT WILL BE EXPECTED FROM YOU:

-INSTALL TWO APPS ON YOUR SMARTPHONE.
-THE FIRST APP WILL TRACK YOUR BEHAVIOUR (THE TIME YOU SPEND OUTSIDE AND INSIDE YOUR HOME, NUMBER AND DURATION OF CALLS, TIME ON THE COMMUNICATION AND SOCIAL APPS - ALL INFO IS SUPERFICIAL, SO WE WILL NOT (AND WILL NOT BE ABLE TO) LOOK INTO WHAT EXACTLY DO YOU DO IN THOSE APPS). THE SECOND APP IS A GAME.
-YOU MUST PLAY THE GAME FOR 5 - 10 MINUTES TWICE A DAY FOR A PERIOD OF 14 DAYS AND FILL A MOOD REPORT BEFORE EACH SESSION.
-YOU WILL ALSO HAVE TO FILL IN QUESTIONNAIRES AT THE START AND THE END OF THE RESEARCH.
-YOU WILL GET A PERSONALIZED REPORT OF YOUR PERFORMANCE AFTER THE STUDY & MORE INFO ON THE STUDY IN GENERAL IF YOU WOULD LIKE TO.

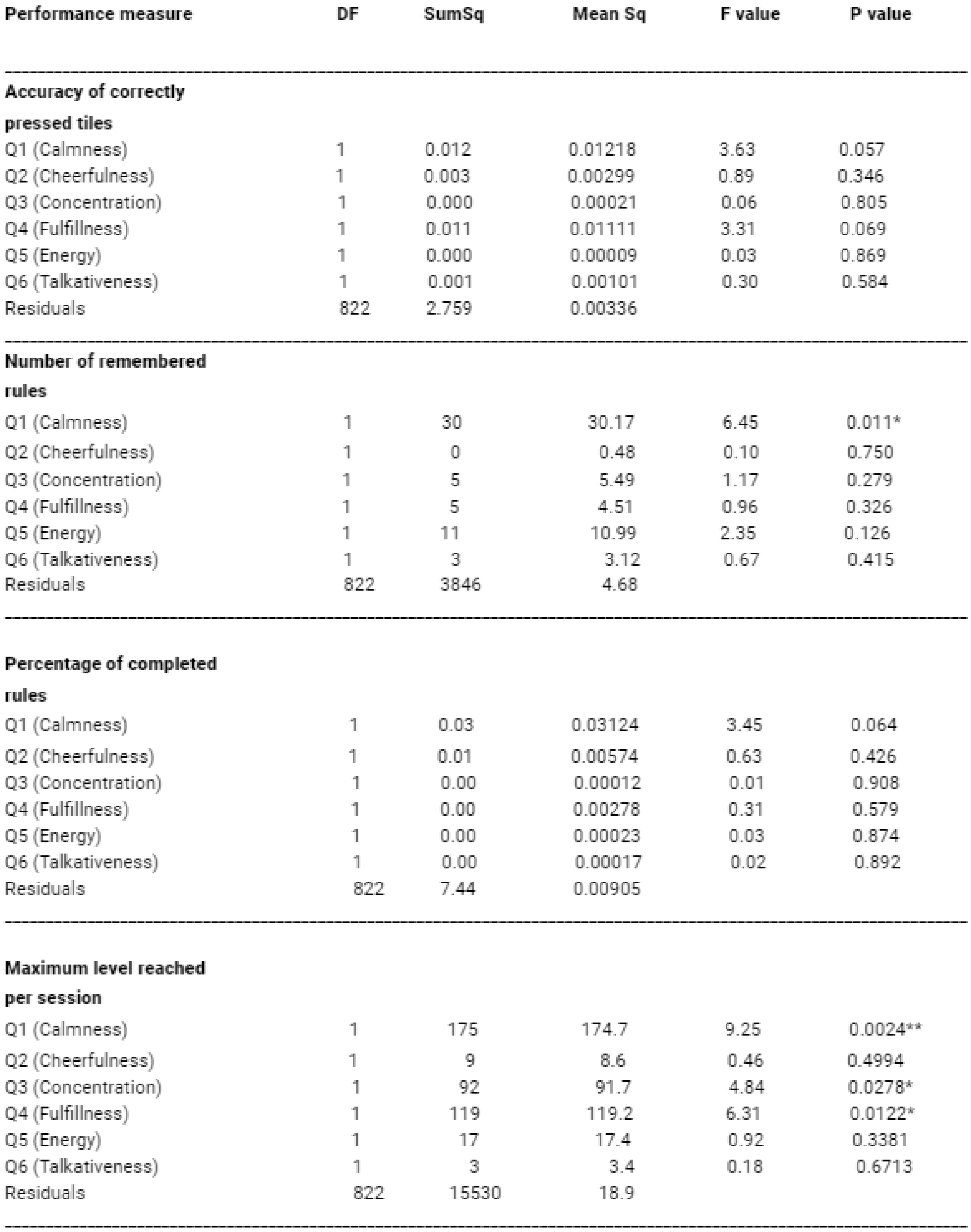
PARTICIPANT REQUIREMENTS:
-YOU MUST HAVE AN ANDROID SMARTPHONE WITH AT LEAST ANDROID VERSION 5.0 (LOLLIPOP)
-AGE SHOULD BE BETWEEN 18-35
-NOT COLOR BLIND
-NOT WORKING LATE NIGHT SHIFTS
-AVAILABLE FOR 14 DAYS IN A ROW WITHOUT EXPECTING LARGE VARIATIONS IN THE DAILY ROUTINE

WHAT YOU GET:
A 20 EURO MONEY COMPENSATION.
PERSONALIZED REPORT ABOUT YOUR BEHAVIOR AND PERFORMANCE.

IF YOU ARE INTERESTED, CONTACT US:
I.TELIATNYKOV@STUDENT.RUG.NL

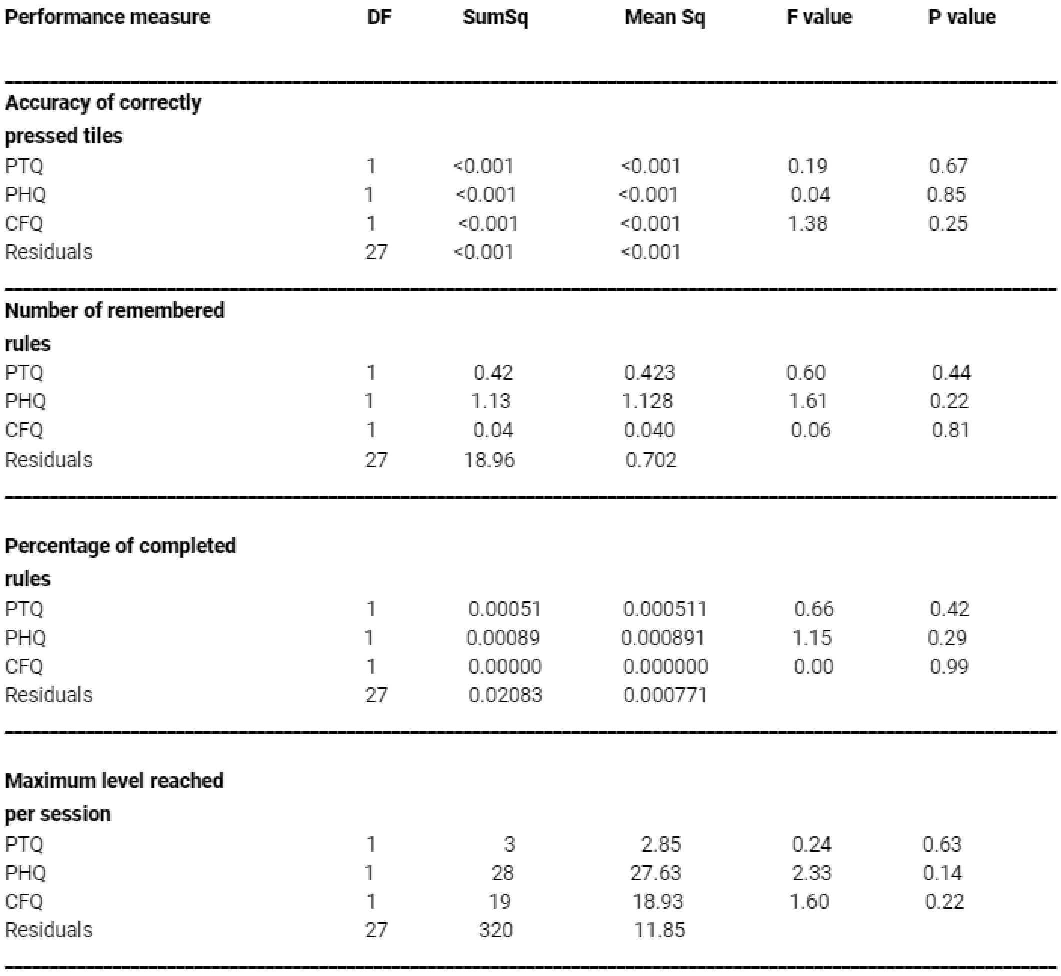
Figure A.12: Information brochure.

B. Appendices



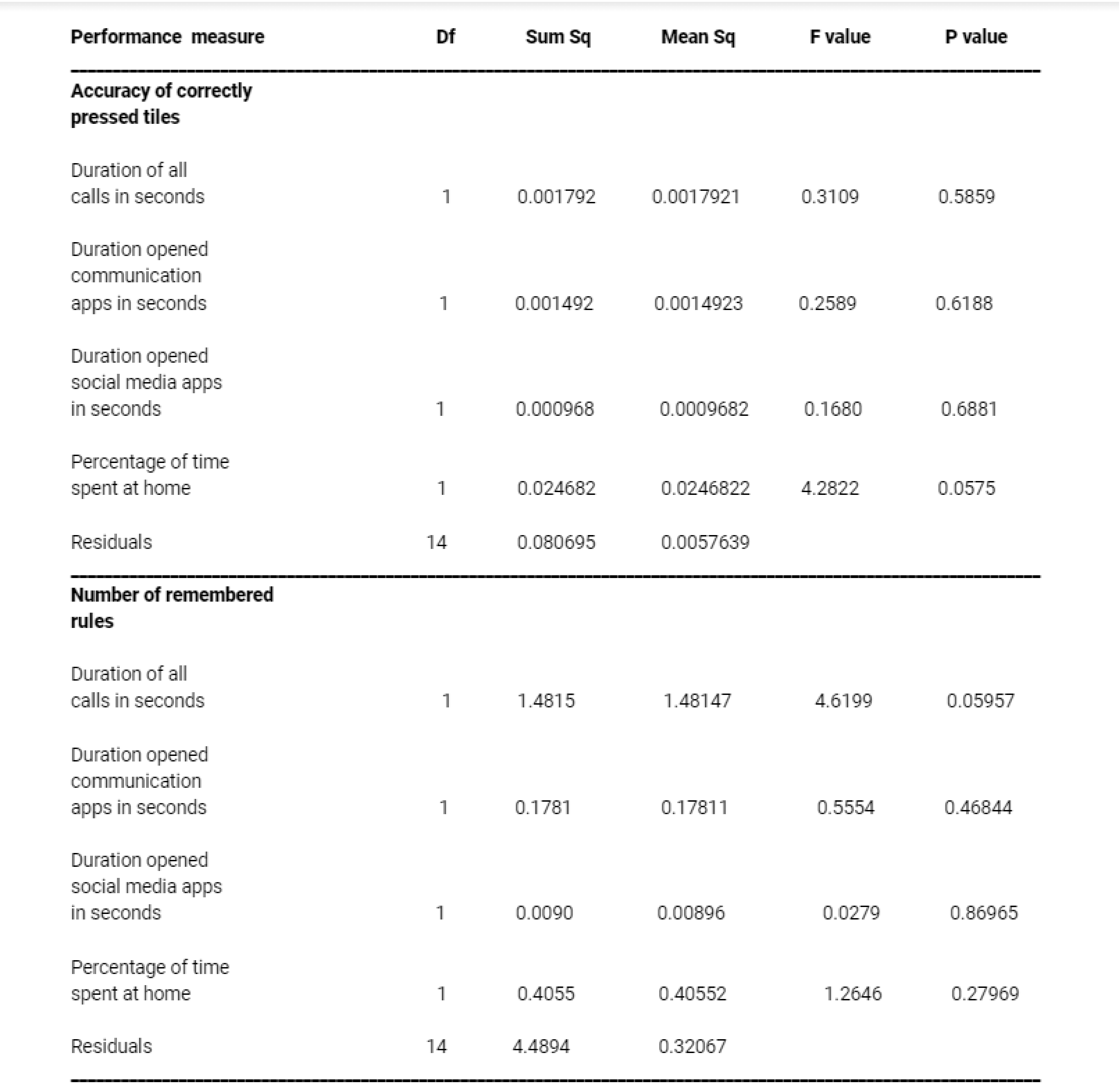
Performance measure	DF	SumSq	Mean Sq	F value	P value
Accuracy of correctly pressed tiles					
Q1 (Calmness)	1	0.012	0.01218	3.63	0.057
Q2 (Cheerfulness)	1	0.003	0.00299	0.89	0.346
Q3 (Concentration)	1	0.000	0.00021	0.06	0.805
Q4 (Fulfillness)	1	0.011	0.01111	3.31	0.069
Q5 (Energy)	1	0.000	0.00009	0.03	0.869
Q6 (Talkativeness)	1	0.001	0.00101	0.30	0.584
Residuals	822	2.759	0.00336		
Number of remembered rules					
Q1 (Calmness)	1	30	30.17	6.45	0.011*
Q2 (Cheerfulness)	1	0	0.48	0.10	0.750
Q3 (Concentration)	1	5	5.49	1.17	0.279
Q4 (Fulfillness)	1	5	4.51	0.96	0.326
Q5 (Energy)	1	11	10.99	2.35	0.126
Q6 (Talkativeness)	1	3	3.12	0.67	0.415
Residuals	822	3846	4.68		
Percentage of completed rules					
Q1 (Calmness)	1	0.03	0.03124	3.45	0.064
Q2 (Cheerfulness)	1	0.01	0.00574	0.63	0.426
Q3 (Concentration)	1	0.00	0.00012	0.01	0.908
Q4 (Fulfillness)	1	0.00	0.00278	0.31	0.579
Q5 (Energy)	1	0.00	0.00023	0.03	0.874
Q6 (Talkativeness)	1	0.00	0.00017	0.02	0.892
Residuals	822	7.44	0.00905		
Maximum level reached per session					
Q1 (Calmness)	1	175	174.7	9.25	0.0024**
Q2 (Cheerfulness)	1	9	8.6	0.46	0.4994
Q3 (Concentration)	1	92	91.7	4.84	0.0278*
Q4 (Fulfillness)	1	119	119.2	6.31	0.0122*
Q5 (Energy)	1	17	17.4	0.92	0.3381
Q6 (Talkativeness)	1	3	3.4	0.18	0.6713
Residuals	822	15530	18.9		

Figure B.1: ANOVA results of the effect of mood questions on game performance



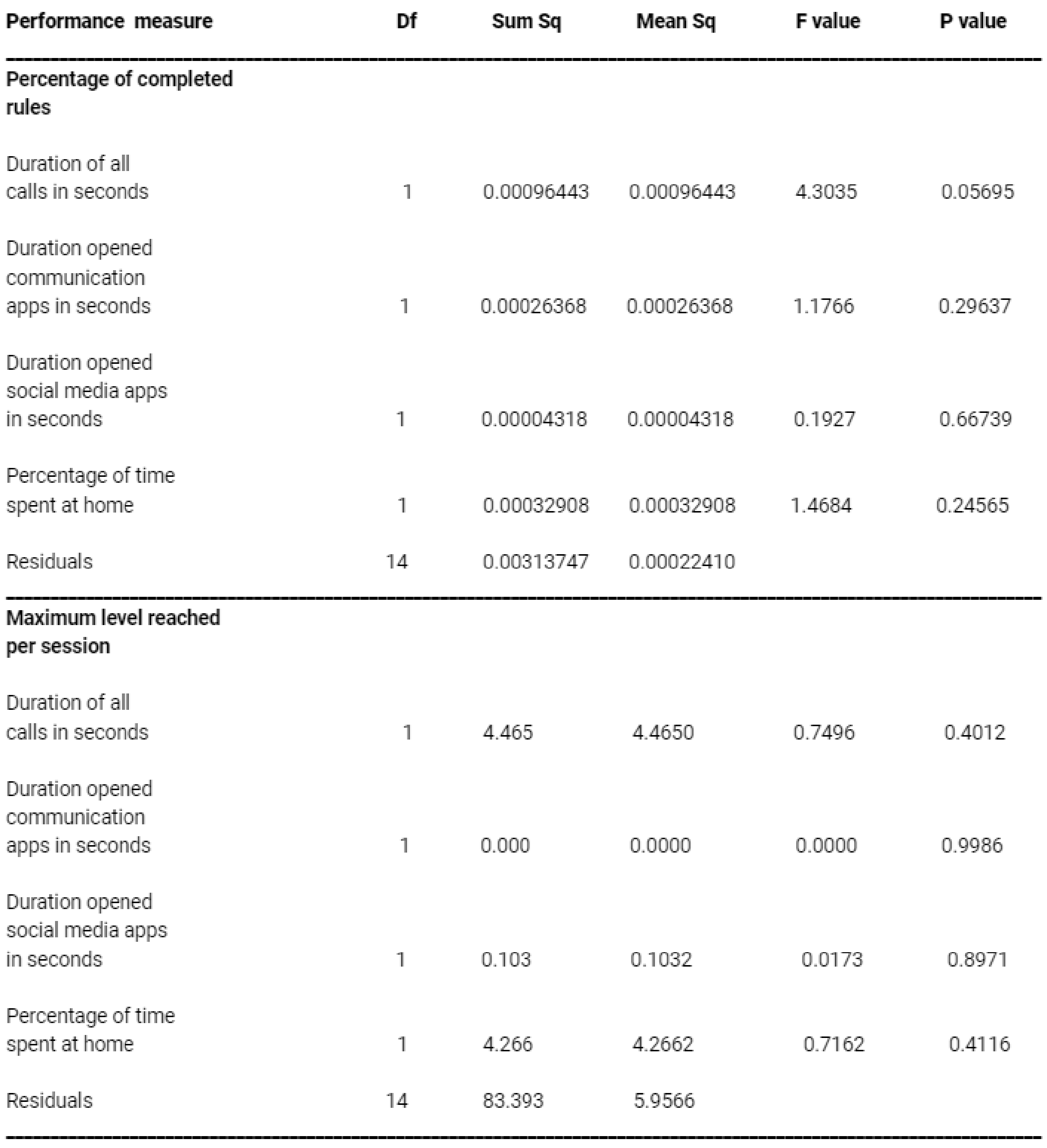
Performance measure	DF	SumSq	Mean Sq	F value	P value
Accuracy of correctly pressed tiles					
PTQ	1	<0.001	<0.001	0.19	0.67
PHQ	1	<0.001	<0.001	0.04	0.85
CFQ	1	<0.001	<0.001	1.38	0.25
Residuals	27	<0.001	<0.001		
Number of remembered rules					
PTQ	1	0.42	0.423	0.60	0.44
PHQ	1	1.13	1.128	1.61	0.22
CFQ	1	0.04	0.040	0.06	0.81
Residuals	27	18.96	0.702		
Percentage of completed rules					
PTQ	1	0.00051	0.000511	0.66	0.42
PHQ	1	0.00089	0.000891	1.15	0.29
CFQ	1	0.00000	0.000000	0.00	0.99
Residuals	27	0.02083	0.000771		
Maximum level reached per session					
PTQ	1	3	2.85	0.24	0.63
PHQ	1	28	27.63	2.33	0.14
CFQ	1	19	18.93	1.60	0.22
Residuals	27	320	11.85		

Figure B.2: The results of ANOVA test on the questionnaires



Performance measure	Df	Sum Sq	Mean Sq	F value	P value
Accuracy of correctly pressed tiles					
Duration of all calls in seconds	1	0.001792	0.0017921	0.3109	0.5859
Duration opened communication apps in seconds	1	0.001492	0.0014923	0.2589	0.6188
Duration opened social media apps in seconds	1	0.000968	0.0009682	0.1680	0.6881
Percentage of time spent at home	1	0.024682	0.0246822	4.2822	0.0575
Residuals	14	0.080695	0.0057639		
Number of remembered rules					
Duration of all calls in seconds	1	1.4815	1.48147	4.6199	0.05957
Duration opened communication apps in seconds	1	0.1781	0.17811	0.5554	0.46844
Duration opened social media apps in seconds	1	0.0090	0.00896	0.0279	0.86965
Percentage of time spent at home	1	0.4055	0.40552	1.2646	0.27969
Residuals	14	4.4894	0.32067		

Figure B.3: ANOVA results of Behapp variables' effect on the accuracy of correctly pressed tiles and on a number of remembered rules



Performance measure	Df	Sum Sq	Mean Sq	F value	P value
Percentage of completed rules					
Duration of all calls in seconds	1	0.00096443	0.00096443	4.3035	0.05695
Duration opened communication apps in seconds	1	0.00026368	0.00026368	1.1766	0.29637
Duration opened social media apps in seconds	1	0.00004318	0.00004318	0.1927	0.66739
Percentage of time spent at home	1	0.00032908	0.00032908	1.4684	0.24565
Residuals	14	0.00313747	0.00022410		
Maximum level reached per session					
Duration of all calls in seconds	1	4.465	4.4650	0.7496	0.4012
Duration opened communication apps in seconds	1	0.000	0.0000	0.0000	0.9986
Duration opened social media apps in seconds	1	0.103	0.1032	0.0173	0.8971
Percentage of time spent at home	1	4.266	4.2662	0.7162	0.4116
Residuals	14	83.393	5.9566		

Figure B.4: ANOVA results of Behapp variables' effect on a percentage of completed rules and on a maximum level reached