university of groningen / faculty of science and engineering

HUMAN-MACHINE COMMUNICATION
DEPARTMENT OF ARTIFICIAL INTELLIGENCE

MASTER'S THESIS
SEPTEMBER, 2020

Tracking Cognition over Time with a Smartphone Game

Patrick T. van der Zwan

Supervisors:
Prof. Dr. N.A. Taatgen
Artificial Intelligence - UNIVERSITY OF GRONINGEN
Dr. M.K. van Vugt
Artificial Intelligence - UNIVERSITY OF GRONINGEN

Abstract

Depression is becoming an increasingly common mental health disease. It can be difficult to notice symptoms of depression, and people often experience several barriers when seeking professional help. The risk of recurrence after diagnosis and treatment is high, and this keeps increasing with every new depressive episode or relapse. Not only is depression a burden on someone's health, but it can take a long time before a relapse in depression is acknowledged, which further delays taking action to seek professional help again. There seems to be a need for a mechanism to monitor the development of symptoms of depression. Since most people have a smartphone which they use everyday, it would be interesting to see whether a smartphone game could provide useful insights in the fluctuations of one's mental health. In this study, we examined the performance on a smartphone game over time while keeping track of the participant's mood and mental health via short mood reports. Participants played a smartphone puzzle game for 14 consecutive days, for two sessions of five minutes each day. A small mood report was administered each session, and a pre-test and post-test were conducted, both consisting of depression-related questionnaires. We found that participants performed better in the smartphone game when they were more concentrated and rated their self-worth higher. Participants remembered more rules and made less errors. In contrast, participants with a higher depression score on the PHQ-9 questionnaire also performed better in the smartphone game: they completed more rules and levels, and remembered more rules. These findings provide interesting possibilities for future research, such as tracking cognition in a clinical sample.

Key words: cognition, depression, mental health, mood, smartphone game

Acknowledgements

I would like to thank my supervisors Niels Taatgen and Marieke van Vugt for their continuous support and trust in this research project, as well as their helpful feedback and guidance. I would also like to thank Fionneke Bos and Marie-José van Tol from the Department of Psychiatry of the University of Groningen, for assisting with the selection of experience sampling questions.

Furthermore, I would like to thank Floor, Arjan, Ymke, Arianne, Stefan and Rebecca for their personal support and assistance, especially for testing the app and giving feedback on the online experiment forms. Finally, I would like to thank my family: my two sisters and my parents, for always having faith in me.

Contents

Chapter 1: Introduction

1.1 Background

Depression is a common mental health disorder. It is estimated that 4.4 percent of the global population is living with depression in 2015 (World Health Organization and others, 2017). It can be difficult to notice whether someone has a depression and is not just feeling sad or down for other (temporary) reasons. Often enough, people in the environment sense that there is something going on, but it can take a long time before the person takes action and gets properly diagnosed by a doctor. Furthermore, people generally hesitate to see a doctor for mental health issues. For example, young people with mental health issues experience several barriers for seeking professional help, such as expensive healthcare, not knowing where to find help, or believing that they could handle their problems on their own (Rickwood et al., 2005). They also perceive stigma and embarrassment or have problems recognising the symptoms (Gulliver et al., 2010).

When a depressed individual eventually visits a doctor and gets diagnosed with depression, they typically receive treatment in the form of Cognitive Therapy (CT) or antidepressant medication. Studies have shown that the chances of experiencing another major depressive episode or relapse are quite high (Boland et al., 2009; Goldman et al., 1999), as more than 75% of diagnosed depressed patients will experience another depressive episode after the first one. After several relapses, this risk is even higher. Relapses are also more frequent when patients stopped using antidepressant medication, compared to patients that had withdrawn from Cognitive Therapy (Hollon et al., 2005). When the symptoms of a relapse are severe enough, one would hopefully visit a doctor again and the cycle of diagnosis and treatment repeats itself. This can take a long time however, as it is difficult for an already depressed or relapsed patient to take initiative and seek professional help once more.

The common way of diagnosing depression starts with looking whether someone experiences symptoms of depression, such as difficulty concentrating, loss of self confidence, sleeping problems, persistent sadness or thoughts about suicide and death (World Health Organization and others, 2017). This can be done with self-report

questionnaires, which contain standard questions that will identify which symptoms are present, in combination with the duration. Using these self-report questionnaires can be quite tedious, and people may choose to not answer truthfully.

Another way to notice depressive symptoms is by evaluating someone's mood over time. This can be done with an external device (Wigman et al., 2015). In this study, participants were given a device to wear on their wrist, which emitted a beep signal at a quasi-random moment for 10 times each day, after which the participants were required to complete a self-assessment diary for several days in a row, to get a measure of fluctuations in mood and mental health. An issue with this study is that self-report questions are often susceptible to a response bias. People have a tendency to report socially desirable responses (Donaldson & Grant-Vallone, 2002), which could result in an under-report of actual depressive symptoms such as negative mood or rumination. More importantly, this method of sampling questions is very invasive. Interrupting someone multiple times a day would not be feasible for longer periods of time. Another study involving self-report used the PHQ-9 questionnaire for depression screening in a mobile app (BinDhim et al., 2016). They found that presenting a questionnaire on a mobile app motivated some users to actively seek a depression diagnosis at a healthcare professional, after the app indicated a moderate or higher depression score.

It would be interesting to explore other options aside from self-report to detect changes in mental health. Smartphones are abundant in our everyday lives. Apart from being able to communicate, users can play games, consume social media and use apps to keep track of their physical well-being. Smartphone games are being used more and more often for studying cognition and mental health. For example, the mobile game Sea Hero Quest collects data on human spatial navigation to learn more about dementia (Morgan, 2016; Spiers et al., 2016). Other apps, such as MoodTrainer provide a virtual Cognitive Behavior Therapy environment which focuses on changing the negative thought process (Addepally & Purkayastha, 2017).

However, apps that monitor or track changes in mental health are less known. A lot of studies focus on capturing smartphone usage data to analyse or predict possible changes in mood. For example, by analysing communication history and app usage, the average daily mood of a user can be inferred with 93% accuracy after 2 months of training on smartphone usage patterns (LiKamWa et al., 2013). Another study that looked at mobile usage patterns found that participants with depression saved fewer contacts on their phones, spent more time on their phone, made less outgoing calls and sent more text messages than participants without depression (Razavi et al., 2020). Doryab et al. (2014) examined sleep patterns, using GPS data to see when people left their homes. They also measured physical activity to determine signs of problems that could lead to depression. One drawback of apps that capture usage data is that they often collect a lot of sensitive information which leads to concerns about data privacy. Most of the people would also criticise that these apps lack proof that they are indeed effective (Lipschitz et al., 2019).

We suggest to use a smartphone app which measures cognitive performance over time while playing a fun game. We will use a modified version of an already existing smartphone game called Wollie. Wollie was originally developed for a study to measure transfer of skills between the smartphone game and multiple cognitive tasks (Doesburg & Taatgen, n.d.). In this puzzle game, images and numbers have to be tapped according to certain rules before the timer runs out. Subjects used Wollie to train their working memory, task-switching and focusing skills. Wollie was developed for Android devices only, using Google's official integrated development environment Android Studio.

We propose an experiment where participants will play the smartphone game Wollie for two weeks, with two sessions of five minutes each day. Wollie will be modified to include a short mood report containing several self-report questions. This way, we will be able to measure fluctuations in mood over time, to see whether these have an effect on the performance in Wollie. Additionally, validated questionnaires will be used at the start and end of the experiment, to inspect individual differences such as depression. Literature suggests that there are numerous cognitive deficits in depression, such as difficulty concentrating (Watts & Sharrock, 1985), memory impairments (Burt et al., 1995) and worse performance on working memory (WM) tasks (Onraedt & Koster, 2014; Rose & Ebmeier, 2006). We are primarily interested in whether these impairments would be reflected in the performance in this smartphone game. We will study this by exploring a novel combination of both self-report questions and fluctuations in performance over time.

1.2 Research Questions

The research questions of this study are:

1. *Does performance in a smartphone game covary with fluctuations in mood?*
2. *Does performance in a smartphone game covary with individual differences in depression?*

The fluctuations in mood will be measured with a short mood report before each playing session in the smartphone game, and the individual differences in depression will be measured with questionnaires before and after the experiment. The measurement of performance in the smartphone game (Wollie) is a little more complex to determine, as Wollie is not a validated laboratory task with one dependent variable such as reaction time. We defined several measures such as error rate, amount of rules completed, levels completed and rules remembered or forgotten. These measurements will be discussed in more detail in the method section 2.5.

We hypothesize that we will find a positive correlation between fluctuations in mood and performance. In other words, we expect that the performance in Wollie will be better when people report positive affect, higher concentration and self-worth. Lower concentration is known to have negative effects on performance (Gaillard, 2008), while sustained negative affect and worthlessness are characteristics of depression (Gotlib & Joormann, 2010), which in turn would have a negative impact on performance (Burt et al., 1995; Onraedt & Koster, 2014; Rose & Ebmeier, 2006; Watts & Sharrock, 1985).

For the individual differences in depression we expect a negative correlation between depression and performance. We hypothesize that people with higher depression scores will perform worse in Wollie, due to the cognitive deficits mentioned earlier.

Chapter 2: Method

2.1 Participants

49 participants completed the full study (27 women, 22 men; mean age 23.3 years, range 18-30 years). These participants received a monetary compensation of 30 euros after completing the full 14-day experiment, including the pre-test and post-test questionnaires. In total, 58 participants (31 women, 27 men; mean age 23.5 years, range 18-31 years) signed up for the study. However, nine participants stopped playing Wollie after several days for various reasons, and consequently did not finish the post-test questionnaires. These participants received a monetary compensation relative to the time they invested until they quit the study. Analysis were conducted using the 49 participants that completed the full study.

2.2 Materials

2.2.1 Mood questions

To measure the fluctuations of mood and mental health over time, we selected four mood or mental health related questions. They were presented in a short mood report at the beginning of each playing session in the smartphone game. These self-report questions were selected through a process of talking to researchers from the Department of Psychiatry, University of Groningen, The Netherlands (M J van Tol 2018, personal communication, 4 June). The final four questions were chosen from a large data bank of experience sampling questions from a crowd-sourcing study that collected self-report data on mental health in a general population sample (Krieke et al., 2016). The first two questions measure positive and negative affect: calmness and cheerfulness, which Krieke et al. (2016) adapted from the PANAS (Watson et al., 1988). The third question measures concentration or focus, and the fourth question measures self-worth or depression. All questions were adapted for daily use, by using the phrase *At this moment I feel [...]*

#	Question
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

Table 2.1: Mood questions used in the short mood report in Wollie.

These four questions all have two versions, a positively formulated one, and a negatively formulated one. Example: *At this moment I feel* ***calm*** (positive) */* ***stressed*** (negative). All eight questions (both positive and negative versions) were used in each short mood report. This was done to ensure a balanced number of positively and negatively formulated questions, to counter a possible acquiescence bias, where people would report positive to all questions regardless of the content of the question (Podsakoff et al., 2003). Using only 8 questions, we are able to capture some essential fluctuations in mood and mental health, without interrupting the participant multiple times a day with long-winded questionnaires.

See Table 2.1 for a quick view of all eight questions, and see Table A.1 for the full table including the Dutch translation. Note that from now on, we will refer to these four questions as *mood questions* for simplicity's sake, although some questions are also mental health related.

2.2.2 Questionnaires

To measure the individual differences in depression/mental health, we used three questionnaires: one that measures depression directly, one that is depression-related and one that measures cognitive deficits or slips of action. These questionnaires were administered in a *pre-test* at the beginning of the experiment, and in a *post-test* at the end of the experiment (after 14 days of playing the smartphone game). The following three questionnaires were used: the Patient Health Questionnaire (PHQ-9), the Perseverative Thinking Questionnaire (PTQ) and the Cognitive Failures Questionnaire (CFQ). See Table B.1, Table B.2 and Table B.3 for the full questionnaires and their Dutch translation.

The PHQ-9 (Löwe et al., 2004) is a shorter version of the full Patient Health Questionnaire, only incorporating the DSM-IV criteria for depression. There are nine

questions in total, and the response scale is a 4-point Likert scale ranging from 0 (not at all) to 3 (nearly every day). Scores on the PHQ-9 range from 0-27. A score from 5-9 is considered as having minimal symptoms, 10-14 as a minor depression or a mild major depression, 15-19 as a major (moderately severe) depression, while scores higher than 20 suggest a severe major depression.

The PTQ (Ehring et al., 2011) is a self-assessment questionnaire to measure someone's general tendency to repetitive negative thinking, which is often one of the symptoms of depression. It consists of 15 items. Answers range from 0 (never) to 4 (always) on a 5-point Likert scale, and scores range from 0-60.

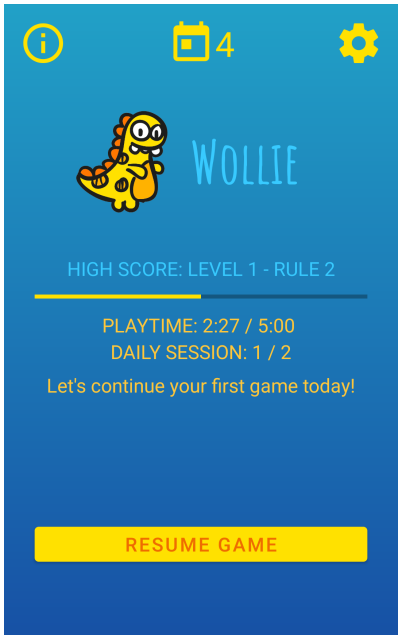
The CFQ (Broadbent et al., 1982) measures self-reported failures in perception, memory and motor function that everyone makes from time to time. There are 25 questions, and answers range from 0 (never) to 4 (very often) on a 5-point Likert scale. Scores range from 0-100.

2.2.3 Wollie

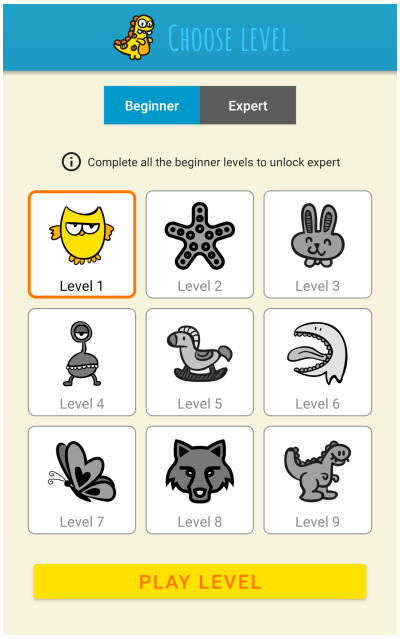
Wollie is a smartphone game originally developed by Inge Doesburg (Doesburg & Taatgen, n.d.). It is based on an existing puzzle game called *RULES!* which is available in the App Store for iOS (www.rulesgame.net). For a list of modifications for Wollie, see Appendix F.

The goal of the game is to empty the board of colorful tiles according to certain rules before the countdown timer runs out. See Figure 2.1 for screenshots of the game. A rule is for example: `tap all green things` or `tap numbers in descending order`. The game gets more difficult as more and more rules need to be followed in succession to clear the board. This requires memorizing the rules and tapping the right tiles before the timer runs out. There are 16 tiles arranged in a 4 by 4 grid, and each tile can be interacted with by tapping on it. All tiles feature a random image and a number between 1-10. See Figure E.1 for all images used in the game. When a tile is correctly pressed, it displays a vanishing animation and is then removed from the board. If a tile is incorrectly pressed (e.g. it does not correspond with the current rule), it flashes red and an error sound is played. Additionally, there is a small delay (0.5 seconds) before a new tile can be pressed, to prevent participants from pressing multiple tiles at once and *guessing* the correct tiles.

Wollie features a total of nine levels in two difficulties: beginner and expert. In beginner difficulty, the timer starts counting down from 30 seconds, and for completing a rule (by clearing the board), 14 seconds are added. In expert difficulty, the countdown timer is set to 20 seconds, and only 9 seconds of extra time is added when a rule is completed. Each level consists of 10 rules to remember, and newer rules are applied first. The first rule is always the same: tap numbers in either *descending* or *ascending* order. The other nine rules are randomly shuffled before starting the level. This is done in order to increase the difficulty of

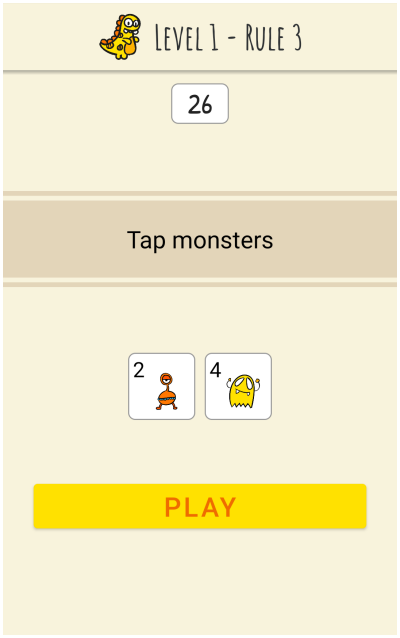


(a) Home screen with progress bar



(b) Choose level screen

(c) Gameplay screen with 16 tiles and a timer



(d) Screen showing which rule to apply next

Figure 2.1: Four screenshots of the smartphone game Wollie.

the game, in which participants ideally reach a performance ceiling. This means that after a few days of playing, we expect that they will not get much better at playing the game. In turn, this would result in a more meaningful comparison between the differences in performance and the day to day fluctuations in mood, instead of just measuring a large learning effect.

A game of Wollie is started by choosing a level. See Figure 2.1b. At first, only level 1 is unlocked and playable. The game begins with the instruction to follow the first rule, for example `tap numbers in descending order`. After tapping all the tiles on the board in time (thus completing rule 1), the timer is paused, and extra seconds are added to the timer depending on the difficulty level (14 seconds in beginner difficulty, and 9 seconds in expert difficulty). Then, the next rule is presented (see Figure 2.1d). For example, rule 2 is: `tap birds`. After pressing the play button, a new board with 16 tiles appears, and the timer starts to count down again. Now tiles with birds have to be tapped first. When all tiles corresponding to this rule are tapped, the game screen flashes yellow, and a notification sound is played, to indicate that a rule has been completed. This means that the previous rule has to be followed. However, this rule is not explicitly stated on the screen: the message only instructs to follow a certain rule *number* (see Figure 2.1c). The previous rule in this example is to follow rule 1 again, which the participant then has to recall from memory. In this case, it was tapping numbers in descending order. A level is completed when all ten rules within that level are completed in time. Participants have to remember up to ten rules, and clear the board of images ten times before the timer runs out. Due to the random shuffling of rules, it would occasionally happen that not all rules have to be applied in order, because some rules cancel each other out. For example: if rule 5 is: `tap all things green`, and rule 3 is: `tap green monsters`, then by completing rule 5, you also completed rule 3, as green monsters would have already been tapped by tapping *all* the green things.

When a level is completed, the next level is unlocked. The difficulty in the game gradually increases with each level, as the rules become more complex and more specific in higher levels. When all nine levels are completed in beginner difficulty, expert difficulty is unlocked.

2.3 Procedure

2.3.1 Online sign-up

Due to the COVID-19 pandemic (Roser et al., 2020), the experiment was fully conducted online. First, informed consent was obtained via Google Forms. In a sign-up form, general information about the research study was explained, including the estimated time needed to complete the study, and the monetary compensation of 30 euros. In addition, criteria for participation were listed:

- 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

An animated image (GIF) showed participants a demo of how to find the Android version on their phones. The informed consent form was shown after the participants have checked all the boxes for the criteria listed above. This online form contained explanations about the purpose of the study, the procedure of the two week period, the monetary reward, possible risks and discomforts (mandatory use of a smartphone everyday), the voluntary nature of participation and the confidentiality of the research data. Contact information was also provided for both the researcher and supervisor.

2.3.2 Pre-test

At the beginning of the pre-test, participants were presented with a language choice: English or Dutch. They were encouraged to only use Dutch when they had difficulty understanding the following words: *ascending, descending, odd, even*. This way, performance in the game should not be affected by people's proficiency in English (when their native language is Dutch).

Participants then had to fill in the three questionnaires as a pre-test (see subsection 2.2.2). After finishing the pre-test, Google accounts were collected to enable participants to download the Wollie app via a closed alpha test-group in the Google Play Store. Participants were guided through the first time setup of Wollie with a separate Google Form. First, they were asked to set two alarms in their default Android alarm clock app each day. This was done to ensure that participants were reminded to play Wollie. The app itself would also try to show a reminder via push notifications. Due to the fact that each Android phone manufacturer has implemented their own battery optimizations, push notifications would unfortunately not work at all for some particular models of smartphones.

A YouTube video was used to guide participants through the installation and setup of Wollie. When opening Wollie for the first time, participants were confronted with a language choice: they were prompted to select the same language they used in the pre-test questionnaires. Next, by pressing the 'get started' button, Android asks permission for Wollie to write to the external storage, which is needed to save the log files of the game. After allowing this, participants were asked to allow Wollie to ignore battery optimizations,

which makes the push notifications more reliable. This message could be ignored if the participant was not able to find the correct setting. Participants then had to fill in their age, sex, and their subject number which they received by e-mail. Subsequently, they were asked to set the Wollie in-app notifications at the same times as their manual alarms in the Android alarm clock app. Participants were informed that they could turn off the manual alarms after a few days, when they could confirm that Wollie's push notifications were working correctly.

The last step in setting up Wollie was to read through the instructions, which briefly summarized what would be expected of the participants for the next 14 days, followed by a short text instruction that explained how to play the game. Participants were able to revisit the instructions screen at any time via the information button on the home screen of the app.

2.3.3 Playing Wollie

The first day of playing was regarded as a practice day, because participants could install the game at any time during the day. Therefore, it was not always possible to play two sessions on the very first day, due to the time restriction between sessions.

Participants had to play two sessions a day for a minimum of five minutes per session. Only the time actually spent playing the game itself counted towards these five minutes, while the time spent in menus or answering short mood reports was ignored. Participants were allowed to play up to ten minutes per session, to prevent ending a session mid-game after exactly five minutes. Participants were restricted to wait at least two hours before starting the second session, after completing the first one. See Figure F.1b. A progress bar and a text message on the home screen of the app indicated how much time was left to play in the current session (see Figure 2.1a). When the minimum play time of five minutes was reached, an exit button would appear, which would save the session and close the app (see Figure F.1a). One session would take place preferably in the morning (after 6:00), and another one in the afternoon or evening (after 15:00). It was impossible to play during the night (between 23:45 and 6:00), because a session that would overlap two different days would be impractical both code and experiment wise.

Sessions that got cancelled mid-game before five minutes of playing time were reached, were discarded and participants had to start over for that particular session. This could happen when participants deliberately pressed the back button to go back to the home screen of the app, or when pausing the app for more than 10 minutes of inactivity.

Each night at 5:00, Wollie would attempt to upload the log files to a secure cloud environment provided by Google Cloud Storage for FireBase. By inspecting special debug log files, it was possible to monitor problems with the reminders or with the game itself during the experiment. Daily insight in the log files also provided information about whether participants were still playing the game, or whether they dropped out.

2.3.4 Post-test

When participants had played Wollie for 14 days, they received a link via e-mail to complete the post-test, which consisted of the same three questionnaires as the pre-test. After completing the post-test, participants were instructed to fill in their personal details to receive the monetary compensation of 30 euros.

2.4 Data analysis

Statistical analyses were carried out using the statistical programming language R (R Core Team, 2020) in RStudio (RStudio Team, 2020). First, the data was analysed and transformed to tidy data using the `tidyr` package (Wickham et al., 2019). Participants that did not complete the full experiment were excluded from the analysis (9 total). These participants reported problems with their smartphones, or underestimated their available time.

To analyse different effects on the performance in Wollie, Linear Mixed Effects (LME) models were fitted using the `lme4` package (Bates et al., 2015), and p-values for the fitted LME models were computed via Satterthwaite's degrees of freedom method using the `lmerTest` package (Kuznetsova et al., 2017). Random intercepts were specified for each participant and each session, thus accounting for the individual differences between subjects, and for differences in performance over time, such as learning effects. Maximum Likelihood (ML) estimation was used for every LME model, to be able to compare models with different fixed effects. ANOVAs were used to compare models with added fixed effects to their base models, to see whether the added degrees of freedom would explain significantly more variance.

2.5 Performance measures

Performance in simple psychological tasks is often measured with one single outcome variable, such as accuracy or response time. However, Wollie is a more complex puzzle game, which requires the participants to react before the timer runs out, memorize rules, switch between rules, and tap the correct tiles corresponding to these rules. To get a good measure of performance, we therefore have to look at multiple variables. We used the following five measures of performance in our analysis:

1. How many levels are completed

2. Percentage of rules completed vs. failed

3. Percentage correct moves vs. incorrect moves (error rate)
4. How many times a rule is forgotten
5. Average number of rules 'remembered' (ranging from 0-10 per level)

Every performance measure is averaged *per session*, because the mood questions are also presented each session.

Levels completed

The number of levels completed could be a good indication of the general performance in Wollie. To complete a level, a participant has to remember up to ten rules and apply them all before the timer runs out. This requires focus and use of memory.

Rules completed

This performance measure is a relative measure: the percentage of rules that are successfully completed. A rule is completed whenever the board of 16 tiles is cleared before the timer has run out. A rule is failed when the timer runs out before all the tiles have been tapped.

Correct moves

Another type of performance is the amount of errors a participant makes. This measure is also relative (a percentage of correct moves). There are two types of moves when a participant taps a tile: correct and incorrect. When a move is correct the tapped tile corresponds with the current rule and disappears from the board. When a move is incorrect the tile will briefly turn red and stay on the board.

Rules forgotten

Tiles can be tapped correctly and incorrectly. An incorrect move can occur due to several reasons, such as overlooking another tile that fits the current rule. Incorrect moves can also be attributed to simple mistakes of tapping the wrong tiles (motor function) or confusion about whether the image corresponds to the rule. However, when a participant presses multiple incorrect tiles in succession, we could reason that they do not remember the current rule anymore. We therefore define that a participant has forgotten a rule when they have tapped more than **three** incorrect tiles in a row. Subsequently, rules forgotten is a measure of the average number of times a rule is forgotten per session.

Rules remembered

Finally, the last performance measure is the average number of rules remembered in a level per session. Each level consists of ten rules. Remembering more rules per level requires good memory and concentration.

2.6 Pre-processing

Mood questions

To analyse the effect of fluctuations in mood on the performance in Wollie, we first have to look at the responses to the mood questions. Each session, four different mood question pairs were presented in a random order. Each question pair consisted of both the positive and the negative formulation, which resulted in eight questions in total. We found that for each question pair, there are moderate to strong negative correlations (polychoric correlation tests with all p's $< .001$). This means that higher responses (stronger agreement) on each positively formulated question are correlated with lower responses (stronger disagreement) on each negatively formulated question, and vice versa. See Table 2.2. We therefore reverse-coded the negatively formulated questions, and then averaged both scores per question, resulting in four questions that were used as a base model for the Linear Mixed Effects model analysis. As these four questions each indicate one concept of measurement, we will refer to them as calmness (q1), cheerfulness (q2), concentration (q3) and self-worth (q4). Note that in every analysis, we started out with this four-question base model, and compared this with another base model where all eight mood questions were added as fixed effects. The eight-question model was only used as a base model, when the added complexity of this model would explain significantly more variance than the model with four questions.

Question	q1a	q1b	q2a	q2b	q3a	q3b	q4a	q4b
At this moment I feel calm (q1a)	1	**-0.56**	0.45	-0.41	0.41	-0.36	0.34	-0.43
At this moment I feel stressed (q1b)	**-0.56**	1	-0.46	0.55	-0.26	0.28	-0.32	0.51
At this moment I feel cheerful (q2a)	0.45	-0.46	1	**-0.65**	0.23	-0.14	0.52	-0.51
At this moment I feel down (q2b)	-0.41	0.55	**-0.65**	1	-0.17	0.22	-0.52	0.64
At this moment I am able to concentrate (q3a)	0.41	-0.26	0.23	-0.17	1	**-0.7**	0.16	-0.13
At this moment I am easily distracted (q3b)	-0.36	0.28	-0.14	0.22	**-0.7**	1	-0.13	0.19
At this moment I feel my life is worth living (q4a)	0.34	-0.32	0.52	-0.52	0.16	-0.13	1	**-0.62**
At this moment I feel I fall short (q4b)	-0.43	0.51	-0.51	0.64	-0.13	0.19	**-0.62**	1

Table 2.2: Polychoric correlation coefficients between the positively and negatively formulated mood questions. Bold numbers are correlations between the question pairs.

Questionnaires

For all the questionnaires, we found no significant difference in scores between the pre-test and the post-test, after performing repeated-measures t-tests and Bayesian t-tests (all p's $> .1$). The time period between the pre-test and post-test was 15 days on average. 40 out of 49 participants (82%) completed the post-test within 15 days of the pre-test, and all participants completed the post-test within 19 days of the pre-test. Because we found no significant difference between the pre-test and the post-test scores, we averaged the pre-test and post-test scores for each of the three questionnaires and used these scores as fixed effects in the Linear Mixed Effects models.

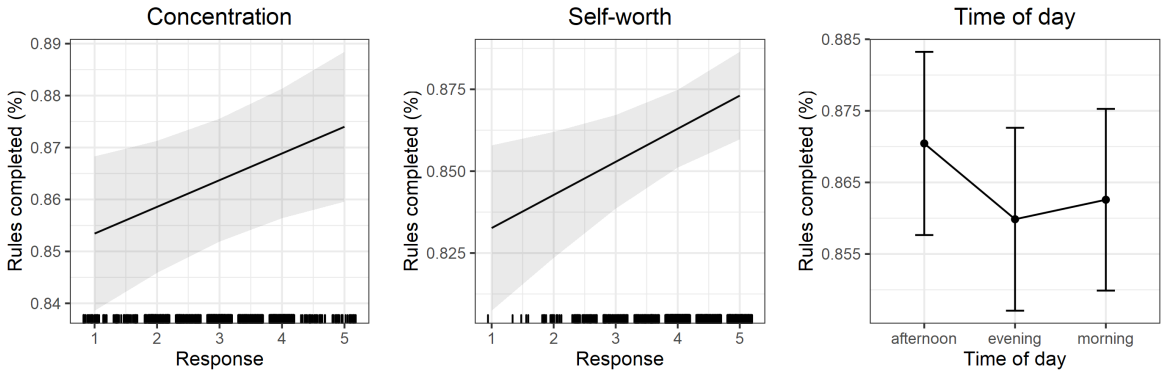
Chapter 3: Results

First, we will present the results of fluctuations in mood over time for each performance measure. After this, we will present the effects of individual differences in depression on the performance in Wollie. We fitted Linear Mixed Effects (LME) models for each performance measure defined in section 2.5. All the tables for the final LME models can be found in Appendix C.

3.1 Fluctuations in mood

Levels completed

The first performance measure is the average number of levels completed per session. The base LME model we used includes participant and subject as random effects, and the four mood questions as fixed effects. Including `time of day` as a fixed effect explains significantly more variance ($\chi^2(2) = 8.10, p < .05$). Furthermore, including the `PHQ-9 score` as a fixed effect also explains significantly more variance ($\chi^2(1) = 11.35, p < .001$). See Table C.1 for the final LME model results. None of the mood questions showed a significant effect. However, a significant effect was found for `PHQ-9 score`, which we will further explain in the individual differences in section 3.2.



(a) **(b)** **(c)**

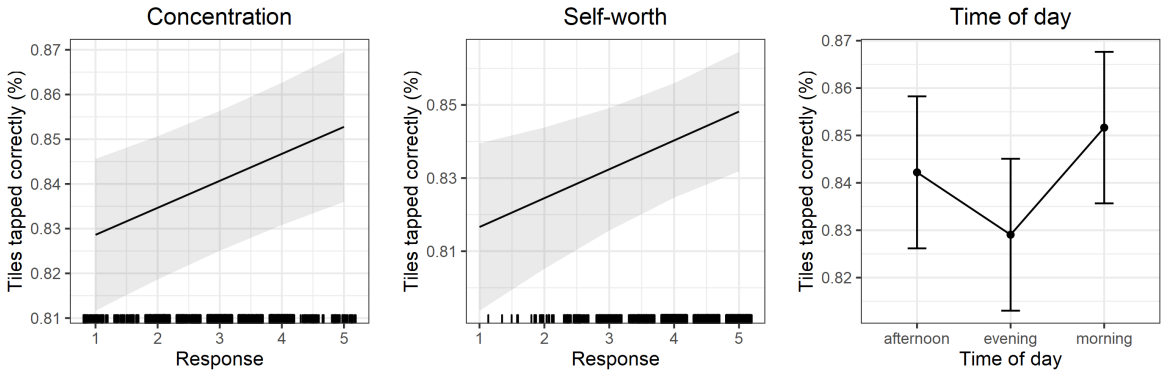
Figure 3.1: Mean percentage of rules completed per session depends on Concentration **(a)**, Self-worth **(b)** and the Time of Day **(c)**. The shaded areas **(a, b)** and error bars **(c)** indicate 95% confidence intervals.

Rules completed

The second performance measure is the average percentage of rules completed per session. For this measure, the base model again includes participant and subject as random effects, and the four mood questions as fixed effects. Including `time of day` as a fixed effect explains significantly more variance ($\chi^2(2) = 6.30, p < .05$) than the model without it. Adding `PHQ-9 score` as another fixed effect explains significantly more variance again ($\chi^2(1) = 8.93, p < .01$). The final model shows a significant effect for `concentration` ($\beta = 0.005, t = 2.34, p < .05$), `self-worth` ($\beta = 0.01, t = 2.78, p < .01$), `time of day: afternoon` ($\beta = 0.01, t = 2.48, p < .05$) and `PHQ-9 score`. The effect of PHQ-9 score will be explained in the individual differences in section 3.2. See Table C.2 for the final LME model results. Looking at Figure 3.1, we can see that participants who reported higher concentration and self-worth completed more rules (relatively). Furthermore, more rules were completed (relatively) in the afternoon (12:00 - 18:00) compared to the evening (18:00 - 0:00) and morning (6:00 - 12:00).

Correct moves

The third performance measure is the average percentage of correct moves, or tiles tapped correctly, per session. The base LME model includes participant and subject as random effects, and the four mood questions as fixed effects. When we include `time of day` as a



(a) **(b)** **(c)**

Figure 3.2: Mean percentage of correctly tapped tiles per session depends on Concentration **(a)**, Self-worth **(b)** and the Time of Day **(c)**. The shaded areas **(a, b)** and error bars **(c)** indicate 95% confidence intervals.

fixed effect, this explains significantly more variance than the model without this variable ($\chi^2(2) = 53.25$, $p < .001$). The final model shows a significant effect for `concentration` ($\beta = 0.006, t = 3.73, p < .001$), `self-worth` ($\beta = 0.007, t = 2.89, p < .01$), `time of day: afternoon` ($\beta = 0.01, t = 4.21$, $p < .001$) and `time of day: morning` ($\beta = 0.02, t = 7.35, p < .001$). See Table C.3 for the final LME model results.

Figure 3.2 shows that participants who reported higher concentration and higher self-worth tapped more tiles correctly (relatively) and, in turn, made less errors when tapping tiles. Also, participants made less errors in the morning (6:00 - 12:00) and in the afternoon (12:00 - 18:00) compared to the evening (18:00 - 0:00).

Rules forgotten

The fourth performance measure is the average number of times a rule is forgotten per session. A rule is defined as forgotten when there are more than three incorrect tiles tapped in succession. The base LME model again has participant and subject as random effects, and the four mood questions as fixed effects. However, when we compared this model with another model which specifies all eight mood questions, the model with eight questions is actually preferred, and explains significantly more variance ($\chi^2(4) = 11.27, p < .05$) than the model with only four mood questions added as fixed effects. We therefore continued with the eight-question-model as a base model. Adding a fixed effect of `time of day` to our base model explains significantly more variance ($\chi^2(2) = 51.55, p < .001$). The final model shows a significant effect for `concentration (q3a)` ($\beta = 0.19, t = 2.01, p < .05$),

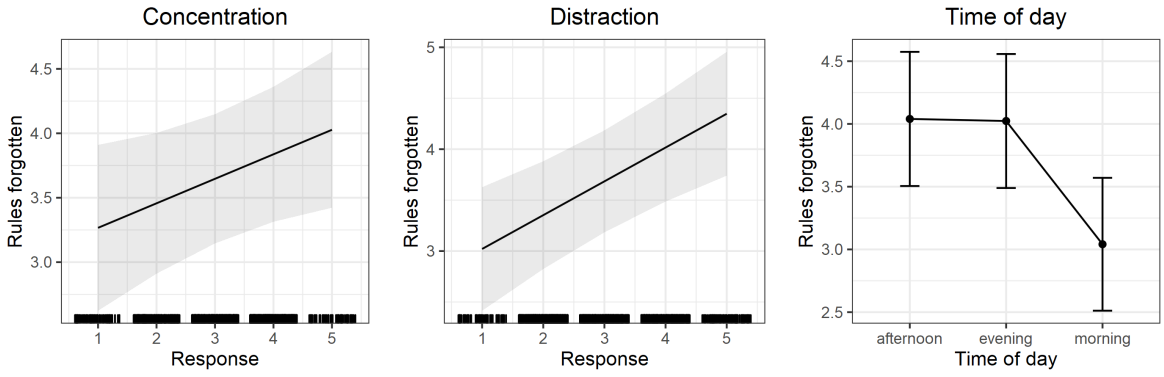


Figure 3.3: Mean number of rules forgotten per session depends on Concentration **(a)**, Distraction **(b)** and the Time of Day **(c)**. The shaded areas **(a, b)** and error bars **(c)** indicate 95% confidence intervals.

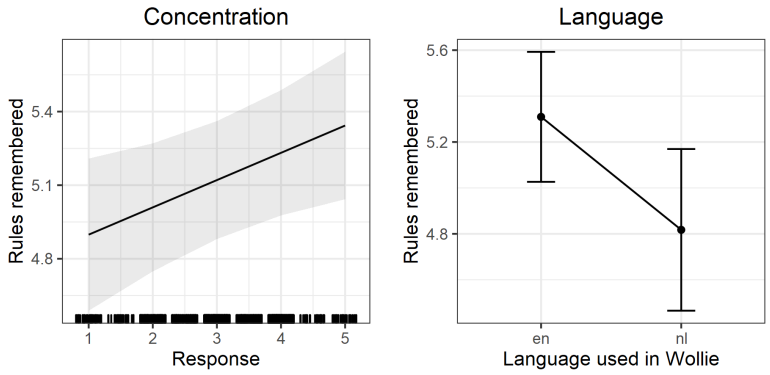
`distraction (q3b)` ($\beta = 0.33, t = 3.79, p < .001$) and `time of day: morning` ($\beta = -0.98, t = -6.13, p < .001$). See Table C.4 for the final LME model results.

In Figure 3.3 we can see a rather contradicting effect: participants forgot more rules on average when they reported higher concentration, but *also* when they reported more distraction. Furthermore, less rules were forgotten in the morning (6:00 - 12:00) compared to the afternoon (12:00 - 18:00) and evening (18:00 - 0:00).

Rules remembered

The fifth and final performance measure is the average number of rules remembered in a level per session. We started with a base LME model with participant and subject as random effects, and the four mood questions as fixed effects. Including `language` as a fixed effect explained significantly more variance ($\chi^2(1) = 6.38, p < .05$). Adding another fixed effect (`PHQ-9 score`) again explains significantly more variance ($\chi^2(1) = 6.51, p < .05$) than the model without this variable. The final model shows a significant effect for `concentration` ($\beta = 0.11, t = 2.32, p < .05$), `language: English` ($\beta = 0.49, t = 2.37, p < .05$) and `PHQ-9 score`. The effect of PHQ-9 score will be explained in the individual differences in section 3.2. See Table C.5 for the final LME model results.

When looking at Figure 3.4, we can see that participants who reported higher concentration remembered more rules on average per level. Furthermore, participants who played Wollie with the app set to English language remembered significantly more rules than participants who played Wollie in Dutch.



(a) **(b)**

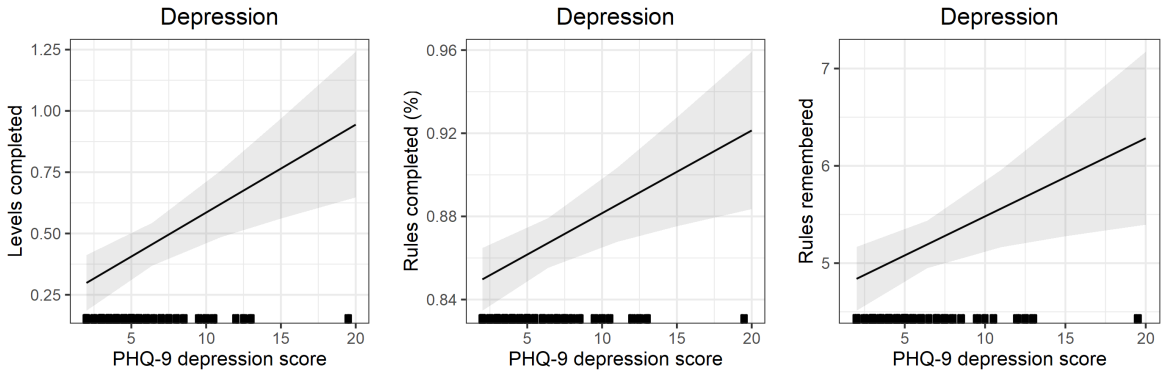
Figure 3.4: Mean number of rules remembered per session depends on Concentration **(a)** and Language **(b)**. The shaded areas indicate 95% confidence intervals.

3.2 Individual differences in depression

We are also interested in whether individual differences in depression between participants could be reflected in the performance in Wollie. These individual differences were measured with three pre-test and post-test questionnaires: the PHQ-9, the PTQ and the CFQ (see subsection 2.2.2 for details about the questionnaires). As explained in section 2.6, we averaged the scores from the pre-test and the post-test to get one score for each questionnaire.

First, we take a look at the performance. Figure 3.5 shows three performance measures where we found an effect of the PHQ-9 depression score. A significant effect is found for levels completed (Figure 3.5a), where participants with a higher PHQ-9 depression score completed more levels on average ($\beta = 0.04, t = 3.54, p < .001$). Furthermore, there is also an effect of rules completed (Figure 3.5b), where participants with a higher PHQ-9 depression score completed a significantly higher percentage of rules ($\beta = 0.004, t = 3.11, p < .01$). Finally, there is an effect of rules remembered (Figure 3.5c), where participants with a higher PHQ-9 depression score also remembered more rules on average ($\beta = 0.08, t = 2.63, p < .05$). No significant effects were found for the other two questionnaires (PTQ and CFQ) on any of the performance measures.

Secondly, it is interesting to look at relations between the reported scores on the questionnaires. As each questionnaire is either measuring depression directly, or depression-related traits, we would expect that higher depression scores on the PHQ-9



(a) **(b)** **(c)**

Figure 3.5: Mean number of levels completed **(a)**, mean number of rules completed **(b)** and mean number of rules remembered **(c)** per session depends on the PHQ-9 depression score. The shaded areas indicate 95% confidence intervals.

Questionnaire	PHQ-9	PTQ	CFQ
Patient Health Questionnaire-9	1	0.65 ***	0.59 ***
Perseverative Thinking Questionnaire	0.65 ***	1	0.43 **
Cognitive Failures Questionnaire	0.59 ***	0.43 **	1

* p <.05 ** p <.01 *** p<.001

Table 3.1: Correlation matrix of scores on the three questionnaires (n = 49).

would also result in higher scores on the PTQ (tendency to repetitive negative thinking). See Table 3.1. There is indeed a strong positive correlation between scores on the PHQ-9 and the PTQ ($r(47) = +.65, p < .001$) Participants with a higher depression score on the PHQ-9 also score higher on the PTQ, indicating that more depressed participants have a higher tendency to repetitive negative thinking. Furthermore, there is also a strong positive correlation between the PHQ-9 and the CFQ ($r(47) = +.59, p < .001$). This indicates that participants with a higher depression score also made more cognitive mistakes. Lastly, there is a moderate positive correlation between the PTQ and the CFQ ($r(47) = +.43, p < .01$). Participants with a higher tendency to repetitive negative thinking also made more cognitive mistakes.

Finally, it is interesting to take a closer look at the individual differences in our sample. When looking at the PHQ-9 scores, 43 out of 49 participants had a score of 9 or lower, indicating minimal symptoms of depression severity. Only 2 out of 49 participants (4%) reported a severe depression score of 20, and 3 out of 49 reported a moderate depression score between 10-14. For the PTQ, we compared our results with student populations in the Netherlands and Belgium, where average PTQ scores of 28.63 (SD = 9.67) and 29.22 (SD = 10.51) were found (Ehring et al., 2012). The PTQ-scores from our study (M = 20.63 for the pre-test, and M = 21.43 for the post-test) are significantly lower on average than those observed in this validation study of the PTQ (Ehring et al., 2012), $t(48) = -4.72, p < .001, d = -0.67$. Apparently, participants in our sample are less concerned with repetitive negative thinking than the average Dutch or Belgian student. It is also interesting to look at 'normal' reported CFQ scores. In a study by Ponds et al. (2006), average CFQ-scores of 30.6 (SD = 10.4) were found in healthy young adults (mean age 30.5, n = 351). However, the CFQ-scores from our study (M = 33.94 for the pre-test, and M = 33.63 for the post-test) are not significantly higher on average than those observed in Ponds et al. (2006), BF = 0.59, $t(48) = 1.70$, $p = .095$, $d = 0.24$.

Chapter 4: Discussion

4.1 Findings

The goal of the present study was to examine whether performance in a smartphone game co-varied with fluctuations in mood and individual differences in depression. We formulated two research questions. The first question was whether performance in a smartphone game covaries with fluctuations in mood. We hypothesized that the performance in Wollie would be better when people would report positive affect, higher concentration, and higher self-worth.

The results partially confirm our hypothesis. Indeed, participants performed better in Wollie when they reported higher *concentration*. A higher percentage of rules was completed, more rules were remembered and more correct moves (less errors) were made when tapping tiles. Participants also performed better when they reported higher *self-worth*: more correct moves were made when tapping tiles (less errors), and a higher percentage of rules was completed.

First off, this is an interesting finding. It may seem trivial that better concentration leads to better performance, as this is in line with the literature (Gaillard, 2008). However, after validating and replicating our study, we would eventually be able to only use the performance measures without the self-report questions to predict fluctuations in someone's concentration. Furthermore, when people rate their self-worth higher, they also perform better. As negative self-worth (worthlessness) is a symptom of depression, the fluctuations in performance in Wollie also seem to be a promising indicator for possible symptoms of depression.

A significant effect of positive affect (calmness or cheerfulness) or negative affect (stressfulness or feeling down) was not found. These two mood questions, based on the PANAS (Watson et al., 1988), did not correlate significantly with any performance measures. As stated in Gotlib and Joormann (2010), sustained negative affect is a core feature of depression. Furthermore, Joormann and Gotlib (2008) found that it is harder for depressed participants to remove irrelevant negative material from working memory. This would suggest that some participants from our study would have performed worse in Wollie when they had responded with more negative affect to our mood questions. One explanation for the fact that we did not find a significant effect is that we may not have had enough

depressed participants in our sample. It could also be the case that playing Wollie is too demanding, which uses up all the cognitive resources of working memory, leaving no room for rumination or repetitive negative thinking.

There is a conflicting finding when we take a look at the performance measure for rules forgotten. Both higher reported concentration *and* distraction correlate with more rules forgotten. We have to take these results with a grain of salt. First off, for this particular performance measure, the model with all eight mood questions (four positively formulated and four negatively formulated) explained significantly more variance than the model with the four questions combined. This resulted in both q3a and q3b (concentration and distraction respectively) having a significant effect. When looking at Figure 3.3, we can also see that the effect is larger for distraction. We are not sure what caused this contradicting effect. It could be the case that these results are skewed by participants who sometimes pressed a lot of tiles in succession, to try and pass a level, for example when they almost remembered all the rules and neared completion of a level. We defined that a rule is 'forgotten' when a participant taps more than three incorrect tiles in a row. It is possible that some participants self-reported that they were able to concentrate, while incorporating frequent tapping of incorrect tiles as part of their strategy to complete more levels. We also have to keep in mind that rules forgotten is a measure that we defined ourselves. It is based solely on the tapping of multiple incorrect tiles, and there could be other unknown reasons why this is happening.

We have to note that the fluctuations we measured with the self-report questions are limited to one or two particular moments per day, when people were reminded to play Wollie. The self-report questions were designed to measure some of the *symptoms* of depression. It may not be representative to relate fluctuations in mood directly with depression. Furthermore, depressed people often do not show symptoms 100% of the time. The mood questions were presented after participants started the app and were ready to play, which could introduce biased responses to the self-report questions. For example, people might have looked forward to playing the game, and responded more positively, even though they could have felt depressed one hour prior to playing.

The second research question of this study was whether performance in a smartphone game covaries with individual differences in depression. For this research question, we hypothesized that people with higher depression scores will perform worse in Wollie, due to several cognitive deficits linked with depression (Burt et al., 1995; Onraedt & Koster, 2014; Rose & Ebmeier, 2006; Watts & Sharrock, 1985). However, the results show a rather opposite effect: people with higher depression scores actually scored *better* in Wollie.

First off, when taking a look at the effect of the depression score on several performance measures (Figure 3.5a, Figure 3.5b and Figure 3.5c), we can see that the size of the confidence

interval increases substantially with higher depression scores. Furthermore, most of the PHQ-9 responses range between 0 and 10. This raises the question whether the effect we found could be attributed to a large influence of a few outliers with a very high PHQ-9 score. We tested this by removing outliers with a depression score larger than 15. Nevertheless, we still found significant results, albeit less strong.

One explanation for this unexpected effect for depression could be that depression does not always appear to show cognitive impairments in every task. Furthermore, Hertel and Rude (1991) found that there was no cognitive deficit for depressed patients when they provided instructions that focused on the task. We could have found a similar effect in Wollie, because participants were repeatedly instructed to follow new rules. Gotlib and Joormann (2010) also summarize that no depressive deficits are found when the participant's attention is controlled by the demands of the task, which eliminates opportunities to ruminate.

However, not only did more depressed participants seem to lack these cognitive deficits, they actually performed *better*. A study by Altamirano et al. (2010) found that mental inflexibility, a trait found in depressed people, could also enhance task performance. Another explanation could be that perfection is also known to be correlated with depression, anxiety and stress (Blankstein et al., 2007; Hewitt & Flett, 1991). Randles et al. (2010) pointed out that perfectionists can respond to feelings of anxiety and self-esteem by displaying workaholic tendencies, which could in turn result in better performance.

An entirely different factor to consider is that the COVID-19 pandemic (Roser et al., 2020) might have had an influence on some people's mental health during the experiment. A lot of things have changed in people's daily routines, and preliminary evidence suggests that an increase of anxiety, depression and stress are common psychological reactions to this pandemic (Rajkumar, 2020; Torales et al., 2020). Only time will tell what the (long-term) effects are on mental health. Several participants reported that playing the smartphone game was one of their highlights of the day during the lockdown. It could be the case that some participants had higher depression scores, but still saw Wollie as a welcoming distraction and tried to perform at their best. This could have resulted in more focused attention and subsequently, a better performance in Wollie.

Finally, we have to keep in mind that our sample is not clinical. There were two severely depressed participants and three moderately depressed participants, when looking at the PHQ-9 results. This means that our statistical power is not very high when looking at depression on its own. It might be possible that there are other effects we just did not find, because they were not significant due to the lack of power.

We found some other significant effects on performance, such as the time of day and the language used in the app. We are not particularly interested in these effects, but they were added as fixed effects to the Linear Mixed Effects models in some cases, because they

explained significantly more variance.

For the time of day effect, we checked whether the amount of sessions played was roughly the same for each time of day segment. This was indeed the case, as the total amount of sessions is 488 in the morning (6:00 - 12:00), 439 in the afternoon (12:00 - 18:00) and 450 in the evening (18:00 - 0:00). First, we found that more rules were completed in the afternoon, compared to the morning and evening (see Figure 3.1c). We also noticed that more correct moves were made in the morning and afternoon, compared to the evening (see Figure 3.2c). These two findings seem to have varying time of day patterns of performance, which is the case for many attention-focused tasks (Kraemer et al., 2000). For some tasks, the peak performance is at noon or in the early afternoon. It is hard to say why this effect occurred without relating this to other measures such as alertness, reported quality of sleep, and so forth.

In addition, less rules were forgotten in the morning (average of 3 per session), compared to the afternoon and evening (4 per session), see Figure 3.3c. This is in line with literature about memory and time of day (Baddeley et al., 1970), where performance on an immediate recall task was better in the morning than in the afternoon.

Lastly, we also found a significant effect of language used in the app on the amount of rules remembered. When playing Wollie, 18 participants used Dutch, while 31 participants used English. The participants that used English remembered more rules on average per level (Figure 3.4b). A possible explanation for this is that the English sentences of each rule are shorter in length than the Dutch translations (see Table D.1), which could make them easier to remember. Note that participants were instructed at the start of the experiment to choose Dutch when they had trouble understanding certain words such as ‘ascending’ and ‘descending’, which could also contribute to this effect.

4.2 Limitations

One limitation of this study is the complication of finding the optimal difficulty of the game. We randomized the order of the rules in each level, which made the levels harder to complete. This was done to prevent finding only a learning effect of participants memorizing the order of each rule. Instead, the goal was to reach a stable performance ceiling. Fluctuations in performance would then probably be caused by other factors such as concentration and depression, which we were interested in.

We expected that people would reach their performance ceiling after a few days of training. This performance ceiling would be different for each person based on their cognitive skills, memory and reaction time. People who would perform better at the game would probably reach the expert difficulty, in which the time was limited even more. We did

not expect that 10 out of the 49 participants would complete all the levels in the expert difficulty, and 47 out of 49 would complete all the levels in the beginner difficulty. As a result, participants who completed all the levels in the expert difficulty were able to play every possible level. This could have made an impact on the performance measures, such as completing more levels or making less errors, due to the fact that these participants were going back to playing easier levels.

Another limitation is that the randomization of the rules also made the levels more unpredictable. This sometimes led to levels being extremely difficult at times, or suddenly much easier, because certain rules would sometimes overlap with each other. For example: when the rule `tap birds` has already been completed, every previous rule containing something to do with a specific bird will be skipped automatically. This could have resulted in more random fluctuations in performance within each session.

Finally, the time that participants spent playing Wollie each session varied between 5 and 10 minutes, which is quite a large amount. We therefore normalized the *levels completed* measurement for each participant by the time that was played in each session, to account for these large variations. Some participants who really liked the game, would play for the maximum amount of time possible, while other participants only played the minimum amount. This resulted in large differences in the total time that participants had played Wollie over the course of the two weeks. To get more uniform playing times, one solution could be to allow the participant to finish their level after the five minute mark is reached, but prevent them from starting a new level. Another solution could be that instead of a time requirement, each subject has a limited number of 'level attempts' each session. For example, every session would consist of a maximum of ten failed attempts (where the timer ran out). This could provide some more reliable performance measures such as how quick each level was solved, the number of levels failed, and how many levels were completed.

Bugs in the app

Not all Wollie sessions were completed by every participant each day, and sometimes sessions would be completed without the answered mood questions. This was due to several bugs. On the first day of the experiment, it became evident that the second session of the day would skip the short mood report. This bug was quickly fixed with an update, but this resulted in some missing data for one session for about half of the total number of participants. Secondly, a rare bug was discovered when a select few participants would e-mail about sessions not being saved. After investigating, this seemed to happen when participants immediately closed the app after playing, by swiping it away from the recent apps list *before* visiting the home screen of the app. If they had exited the app via the exit button on the home screen, the sessions would have been saved correctly.

4.3 Future work

A possible follow-up study would include a control group to further validate the performance measures of Wollie. One group would play Wollie, while the control group would play a validated neuropsychological task, such as the Wisconsin Card Sorting Test (WCST). Multiple studies have shown that depressive patients demonstrate cognitive deficits in this task (Grant et al., 2001; Martin et al., 1991; Merriam et al., 1999). Furthermore, people with higher ruminative tendencies would also make more perseverance errors (Davis & Nolen-Hoeksema, 2000). It would be interesting to see whether the same relations between fluctuations in mood and performance could be found in the control group who would play the WCST.

Another option for future work is to revise the mood questions, and replace some questions with more rumination-focused or cognitive failure related questions. We can see from the questionnaire results that there were strong positive correlations between the PHQ-9 and the other two questionnaires, which further indicates that more depressed people engage in more repetitive negative thinking and make more cognitive mistakes. Although we found no significant results from the PTQ or the CFQ on performance yet, this could be interesting to measure as a fluctuation over time.

Wollie itself could be subject to improvement. A suggestion would be to continue testing the app in multiple pilot studies. This was not possible in the current study due to time constraints. An example of a possible improvement is the difficulty of the game. We received reports that people found the game frustrating to play, when they could not pass a certain level for multiple days in a row. In a future study in which Wollie would be used, this should be taken into consideration. In addition, it was technically possible to 'cheat' Wollie by writing the rules on a piece of paper during the presentation of each new rule, because there was no time constraint on pausing the game when a new rule was introduced. We instructed participants to play fair and only use their smartphone. However, limiting the time the participant can wait before continuing with the newly presented rule would be a better solution.

Eventually, a clinical study with diagnosed depressed patients would be very interesting to conduct. With a larger sample of depressed participants, it might be possible to find more effects of certain fluctuations of mood on various performance measures. Ideally, such a study would take a much longer time (for example three to six months), and the mood questions would be removed from the app, to ultimately see whether a possible relapse could be predicted by fluctuations in the relevant performance measures in Wollie.

Chapter 5: Conclusion

This study has aimed to find whether performance in a smartphone game covaried with fluctuations in mood and individual differences in depression. We have found that better performance in the smartphone game Wollie positively correlates with higher reported concentration and self-worth. In addition, higher performance was also observed in participants with higher depression scores, which was not what we expected.

First, we have shown that there is a significant correlation between concentration and performance, and self-worth and performance. Better performance was observed in the amount of rules completed, rules remembered and correct tiles tapped when concentration was higher. In addition, performance was also better when looking at the percentage of rules completed and the correct tiles tapped when self-worth was rated higher.

We can conclude that these performance measures in Wollie could be a promising way to track fluctuations in people's concentration and self-worth over time. As both concentration and self-worth are indications of symptoms of depression, this could eventually be used in a study to monitor depressed or relapsed patients.

On the other hand, when looking at individual differences the PHQ-9 depression scores suggest that more depressed people perform better in Wollie. We cannot directly compare these results, as one is based on an average score from a questionnaire and the other is based on fluctuations in mood over time, which was assessed with short self-report questions. More research is needed to investigate the relation between the PHQ-9 score and performance. We therefore have to keep in mind that depression does not always seem to show cognitive impairments in every task. Traits such as perfectionism or mental inflexibility could even contribute to more focus and better performance.

To conclude, this study is explorative in nature. We had a lot of freedom to gather information about which performance measures to use, and which questions to use to measure the fluctuations in mood over time. With this in mind, more research is needed to validate the results we found. A good start would be to replicate this study, with possibly a modification such as adding a control group with a validated neuropsychological task. It would also be interesting to investigate whether the conflicting effect of PHQ-9 score on performance would be present in another study. In addition, our nonclinical sample only

consisted of a few depressed participants. In future experiments, the fluctuations in performance in Wollie could be studied in a clinical sample with actual depressed patients. Ultimately, we need to have a better understanding whether these performance fluctuations over time could be used to monitor or even predict symptoms of depression or relapses.

References

Addepally, S. A., & Purkayastha, S. (2017). Mobile-application based cognitive behavior therapy (cbt) for identifying and managing depression and anxiety, In *International conference on digital human modeling and applications in health, safety, ergonomics and risk management*. Springer.

Altamirano, L. J., Miyake, A., & Whitmer, A. J. (2010). When mental inflexibility facilitates executive control: Beneficial side effects of ruminative tendencies on goal maintenance. *Psychological Science*, *21*(10), 1377–1382.

Baddeley, A., Hatter, J., Scott, D., & Snashall, A. (1970). Memory and time of day. *The Quarterly Journal of Experimental Psychology*, *22*(4), 605–609.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48.

BinDhim, N. F., Alanazi, E. M., Aljadhey, H., Basyouni, M. H., Kowalski, S. R., Pont, L. G., Shaman, A. M., Trevena, L., & Alhawassi, T. M. (2016). Does a mobile phone depression-screening app motivate mobile phone users with high depressive symptoms to seek a health care professional's help? *Journal of medical Internet research*, *18*(6), e156.

Blankstein, K. R., Lumley, C. H., & Crawford, A. (2007). Perfectionism, hopelessness, and suicide ideation: Revisions to diathesis-stress and specific vulnerability models. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *25*(4), 279–319.

Boland, R. J., Keller, M. B., Gotlib, I., & Hammen, C. (2009). Course and outcome of depression. *Handbook of depression*, *2*, 23–43.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The cognitive failures questionnaire and its correlates. *British journal of clinical psychology*, *21*(1), 1–16.

Burt, D. B., Zembar, M. J., & Niederehe, G. (1995). Depression and memory impairment: A meta-analysis of the association, its pattern, and specificity. *Psychological bulletin*, *117*(2), 285.

Davis, R. N., & Nolen-Hoeksema, S. (2000). Cognitive inflexibility among ruminators and nonruminators. *Cognitive therapy and research*, *24*(6), 699–711.

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

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of business and Psychology*, *17*(2), 245–260.

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

Ehring, T., Raes, F., Weidacker, K., & Emmelkamp, P. M. G. (2012). Validation of the dutch version of the perseverative thinking questionnaire (PTQ-NL). *European Journal of Psychological Assessment*, *28*(2), 102–108.

Ehring, T., Zetsche, U., Weidacker, K., Wahl, K., Schönfeld, S., & Ehlers, A. (2011). The perseverative thinking questionnaire (ptq): Validation of a content-independent measure of repetitive negative thinking. *Journal of behavior therapy and experimental psychiatry*, *42*(2), 225–232.

Gaillard, A. W. (2008). Concentration, stress and performance. *Performance under stress*, 59–75.

Goldman, L. S., Nielsen, N. H., Champion, H. C., & Council on Scientific Affairs, A. M. A. (1999). Awareness, diagnosis, and treatment of depression. *Journal of general internal medicine*, *14*(9), 569–580.

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual review of clinical psychology*, *6*, 285–312.

Grant, M. M., Thase, M. E., & Sweeney, J. A. (2001). Cognitive disturbance in outpatient depressed younger adults: Evidence of modest impairment. *Biological psychiatry*, *50*(1), 35–43.

Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: A systematic review. *BMC psychiatry*, *10*(1), 113.

Hertel, P. T., & Rude, S. S. (1991). Depressive deficits in memory: Focusing attention improves subsequent recall. *Journal of Experimental Psychology*, *120*(3), 301.

Hewitt, P. L., & Flett, G. L. (1991). Dimensions of perfectionism in unipolar depression. *Journal of abnormal psychology*, *100*(1), 98.

Hollon, S. D., DeRubeis, R. J., Shelton, R. C., Amsterdam, J. D., Salomon, R. M., O'Reardon, J. P., Lovett, M. L., Young, P. R., Haman, K. L., Freeman, B. B., Et al. (2005). Prevention of relapse following cognitive therapy vs medications in moderate to severe depression. *Archives of general psychiatry*, *62*(4), 417–422.

Joormann, J., & Gotlib, I. H. (2008). Updating the contents of working memory in depression: Interference from irrelevant negative material. *Journal of abnormal psychology*, *117*(1), 182.

Kraemer, S., Danker-Hopfe, H., Dorn, H., Schmidt, A., Ehlert, I., & Herrmann, W. M. (2000). Time-of-day variations of indicators of attention: Performance, physiologic parameters, and self-assessment of sleepiness. *Biological psychiatry*, *48*(11), 1069–1080.

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.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., Et al. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of statistical software*, *82*(13), 1–26.

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*.

Lipschitz, J., Miller, C. J., Hogan, T. P., Burdick, K. E., Lippin-Foster, R., Simon, S. R., & Burgess, J. (2019). Adoption of mobile apps for depression and anxiety: Cross-sectional survey study on patient interest and barriers to engagement. *JMIR mental health*, *6*(1), e11334.

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, 1194–1201.

Martin, D. J., Oren, Z., & Boone, K. (1991). Major depressives' and dysthymics' performance on the wisconsin card sorting test. *Journal of clinical psychology*, *47*(5), 684–690.

Merriam, E. P., Thase, M. E., Haas, G. L., Keshavan, M. S., & Sweeney, J. A. (1999). Prefrontal cortical dysfunction in depression determined by wisconsin card sorting test performance. *American Journal of Psychiatry*, *156*(5), 780–782.

Morgan, J. (2016). Gaming for dementia research: A quest to save the brain. *The Lancet Neurology*, *15*(13), 1313.

Onraedt, T., & Koster, E. H. W. (2014). Training working memory to reduce rumination. *PLoS ONE*, *9*(3), e90632.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Ponds, R., Boxtel, M., & Jolles, J. (2006). De 'cognitive failure questionnaire' als maat voor subjectief cognitief functioneren. *Tijdschrift voor Neuropsychologie*, *1*, 37–45.

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.2). Vienna, Austria. https://www.R-project.org/

Rajkumar, R. P. (2020). Covid-19 and mental health: A review of the existing literature. *Asian journal of psychiatry*, 102066.

Randles, D., Flett, G. L., Nash, K. A., McGregor, I. D., & Hewitt, P. L. (2010). Dimensions of perfectionism, behavioral inhibition, and rumination. *Personality and Individual Differences*, *49*(2), 83–87.

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.

Rickwood, D., Deane, F. P., Wilson, C. J., & Ciarrochi, J. (2005). Young people's help-seeking for mental health problems. *Australian e-journal for the Advancement of Mental health*, *4*(3), 218–251.

Rose, E. J., & Ebmeier, K. (2006). Pattern of impaired working memory during major depression. *Journal of affective disorders*, *90*(2-3), 149–161.

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus pandemic (covid-19). *Our World in Data*. https://ourworldindata.org/coronavirus

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Spiers, H., Hornberger, M., Bohbot, V., Dalton, R., Hölscher, C., Manley, E., Sami, S., Silva, R., & Weiner, J. (2016). Sea hero quest. *London: Glitchers*. www.seaheroquest.com/

Torales, J., O'Higgins, M., Castaldelli-Maia, J. M., & Ventriglio, A. (2020). The outbreak of covid-19 coronavirus and its impact on global mental health. *International Journal of Social Psychiatry*, 0020764020915212.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of personality and social psychology*, *54*(6), 1063.

Watts, F. N., & Sharrock, R. (1985). Description and measurement of concentration problems in depressed patients. *Psychological Medicine*, *15*(2), 317–326.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wigman, J. T. W., Van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., Viechtbauer, W., Myin-Germeys, I., & Wichers, M. (2015). Exploring the underlying structure of mental disorders: Cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychological medicine*, *45*(11), 2375–2387.

World Health Organization and others. (2017). *Depression and other common mental disorders: Global health estimates* (tech. rep.). World Health Organization.

Appendix A: Mood questions

Question	English	Dutch translation	Response range	Range	Origin
1a	At this moment I feel calm	Op dit moment voel ik me kalm			PANAS Positive affect
1b	At this moment I feel stressed	Op dit moment voel ik me gestresst			PANAS Negative affect
2a	At this moment I feel cheerful	Op dit moment voel ik me opgewekt	”Strongly disagree”,		PANAS Positive affect
2b	At this moment I feel down	Op dit moment voel ik me somber	”Disagree”,		PANAS Negative affect
3a	At this moment I am able to concentrate	Op dit moment kan ik me goed concentreren	”Neutral”,	1-5	
3b	At this moment I am easily distracted	Op dit moment ben ik snel afgeleid	”Agree”,		Concentration/decision making
4a	At this moment I feel my life is worth living	Op dit moment vind ik mijn leven de moeite waard	”Strongly agree”		
4b	At this moment I feel I fall short	Op dit moment heb ik het gevoel tekort te schieten			Depression/self-worth

Table A.1: The mood questions used at the start of each Wollie playing session.

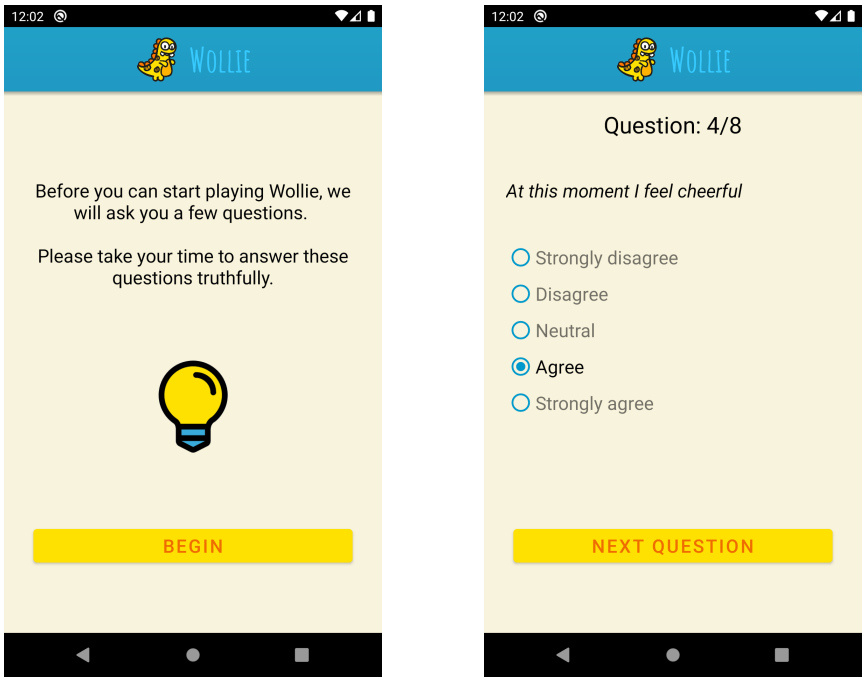


Figure A.1: Screenshots of the short mood report screens in Wollie.

Appendix B: Questionnaires

Question	Dutch translation	Response range	Range
Over the past 2 weeks, how often have you been bothered by any of the following problems?	*Hoe vaak hebt u in de afgelopen 2 weken last gehad van één of meer van de volgende problemen?*		
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.	1. Weinig interesse of plezier in activiteiten 2. Zich neerslachtig, depressief of hopeloos voelen 3. Moeilijk inslapen, moeilijk doorslapen of te veel slapen 4. Zich moe voelen of gebrek aan energie hebben 5. Weinig eetlust of overmatig eten 6. Een slecht gevoel hebben over uzelf - of het gevoel hebben dat u een mislukkeling bent of het gevoel dat u zichzelf of uw familie teleurgesteld hebt 7. Problemen om u te concentreren, bijvoorbeeld om de krant te lezen of om tv te kijken 8. Zo traag bewegen of zo langzaam spreken dat andere mensen dit opgemerkt kunnen hebben? Of het tegenovergestelde, zo zenuwachtig of rusteloos zijn dat u veel meer bewoog dan gebruikelijk 9. De gedachte dat u beter dood zou kunnen zijn of de gedachte uzelf op een bepaalde manier pijn te doen	"Not at all", "Several days", "More than half the days", "Nearly every day"	0-3
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?	10. Als u enig probleem hebt ingevuld, hoe moeilijk maakten deze problemen het dan voor u om uw werk of uw taken in en om het huis te doen, of om met andere mensen om te gaan?	"Not difficult at all", "Somewhat difficult", "Very difficult", "Extremely difficult"	

Table B.1: The Patient Health Questionnaire (PHQ-9).

Question	Dutch translation	Response range	Range
Please read the following statements and rate the extent to which they apply to you when you think about negative experiences or problems.	*Lees de onderstaande uitspraken door en geef aan in welke mate ze voor u van toepassing zijn wanneer u over negatieve ervaringen of problemen nadenkt.*		
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.	1. Dezelfde gedachten blijven steeds door mijn hoofd gaan. 2. Mijn gedachten dringen zich aan mij op. 3. Ik kan niet ophouden om erover na te denken. 4. Ik denk over veel problemen na zonder ze op te lossen. 5. Als ik over mijn problemen nadenk, kan ik op dat moment niets anders doen. 6. Mijn gedachten herhalen zich. 7. Mijn gedachten komen in mij op zonder dat ik dat wil. 8. Ik loop bij bepaalde onderwerpen vast en kan ze moeilijk loslaten. 9. Ik blijf mijzelf vragen stellen zonder een antwoord te vinden. 10. Mijn gedachten belemmeren me om me op andere dingen te richten. 11. Ik blijf de hele tijd over hetzelfde nadenken. 12. Mijn gedachten komen vanzelf in me op. 13. Ik voel me gedwongen om steeds over hetzelfde na te blijven denken. 14. Mijn gedachten helpen mij niet veel verder. 15. Mijn gedachten nemen mij volledig in beslag.	"Never", "Rarely", "Sometimes", "Often", "Always"	0-4

Table B.2: The Perseverative Thinking Questionnaire (PTQ).

Question	**Dutch translation**	**Response range**	**Range**
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.	*De volgende vragen gaan over kleine, alledaagse vergissingen die iedereen van tijd tot tijd maakt. Sommige van die vergissingen overkomen u waarschijnlijk wat vaker dan andere. We willen weten hoe vaak deze dingen u de afgelopen 6 maanden zijn overkomen.*		
1. Do you read something and find you haven't been thinking about it and must read it again?	1. Iets lezen en vlak daarna niet meer weten wat u nu gelezen hebt, zodat u het moet overlezen		
2. Do you find you forget why you went from one part of the house to the other?	2. Vergeten waarom u naar een bepaald gedeelte van uw huis bent gelopen		
3. Do you fail to notice signposts on the road?	3. Wegwijzers over het hoofd zien		
4. Do you find you confuse right and left when giving directions?	4. Links en rechts verwarren bij het beschrijven van een route		
5. Do you bump into people?	5. Per ongeluk tegen mensen opbotsen		
6. Do you find you forget whether you've turned off a light or a fire or locked the door?	6. Niet meer weten of u het licht of het gas hebt uitgedaan, of de deur hebt afgesloten		
7. Do you fail to listen to people's names when you are meeting them?	7. Niet luisteren naar de naam van een persoon op het moment dat deze persoon zich aan u voorstelt		
8. Do you say something and realize afterwards that it might be taken as insulting?	8. Iets er uitflappen en achteraf bedenken dat dat wel eens beledigend voor iemand zou kunnen zijn		
9. Do you fail to hear people speaking to you when you are doing something else?	9. Niet merken dat iemand iets tegen u zegt als u met iets anders bezig bent	"Never",	
10. Do you lose your temper and regret it?	10. Boos worden en daar later spijt van hebben	"Very rarely",	
11. Do you leave important letters unanswered for days?	11. Belangrijke brieven dagenlang onbeantwoord laten	"Occasionally",	0–4
12. Do you find you forget which way to turn on a road you know well but rarely use?	12. Vergeten welke straat u moet inslaan als u een route kiest die u goed kent, maar die u maar zelden gebruikt	"Quite often",	
13. Do you fail to see what you want in a supermarket (although it's there)?	13. In een supermarkt niet kunnen vinden wat u zoekt terwijl het er wel is	"Very often"	
14. Do you find yourself suddenly wondering whether you've used a word correctly?	14. U plotseling afvragen of u een woord op de juiste manier gebruikt		
15. Do you have trouble making up your mind?	15. Moeite hebben met het nemen van een beslissing		
16. Do you find you forget appointments?	16. Afspraken vergeten		
17. Do you forget where you put something like a newspaper or a book?	17. Vergeten waar u iets hebt neergelegd, zoals een boek of een krant		
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?	18. Per ongeluk iets weggooien dat u nodig hebt en bewaren wat u weg wilde gooien		
19. Do you daydream when you ought to be listening to something?	19. Dagdromen terwijl u eigenlijk naar iets of iemand zou moeten luisteren		
20. Do you find you forget people's names?	20. Namen van mensen vergeten		
21. Do you start doing one thing at home and get distracted into doing something else (unintentionally)?	21. Beginnen met iets maar het niet afmaken, omdat u ongemerkt met iets anders bent begonnen		
22. Do you find you can't quite remember something although it's "on the tip of your tongue"?	22. Niet op een woord kunnen komen terwijl het 'op het puntje van uw tong' ligt		
23. Do you find you forget what you came to the shops to buy?	23. In een winkel vergeten wat u kwam kopen		
24. Do you drop things?	24. Dingen uit uw handen laten vallen		
25. Do you find you can't think of anything to say?	25. In een gesprek niets meer weten om over te praten		

Table B.3: The Cognitive Failures Questionnaire (CFQ).

Appendix C: Linear Mixed Effects models results

Model: *LevelsCompleted_norm* $\sim$ *q1 + q2 + q3 + q4 + Time_of_day +* PHQ_Score + *(1|Subject) + (1|SessionID)*

Predictor	**Estimate**	**std. Error**	**t value**	$Pr(>\|z\|)$	
Intercept	0.041	0.154	0.264	0.792	
Calmness (q1)	0.015	0.025	0.593	0.553	
Cheerfulness (q2)	-0.006	0.026	-0.219	0.827	
Concentration (q3)	0.037	0.020	1.866	0.062	
Self-worth (q4)	0.007	0.032	0.227	0.820	
Time of day [afternoon]	0.074	0.038	1.912	0.056	
Time of day [morning]	-0.036	0.038	-0.934	0.350	
PHQ-9 depression score	0.036	0.010	3.541	<0.001	***

* p < .05 ** p < .01 *** p < .001

Table C.1: Results of the linear mixed effect model using the *Mood questions*, the *Time of Day* and the *PHQ-9 score* to predict the average number of **Levels Completed** per session.

Model: *RulesCompletedPct* $\sim$ *q1 + q2 + q3 + q4 + Time_of_day +* PHQ_Score + *(1|Subject) + (1|SessionID)*

Predictor	**Estimate**	**std. Error**	**t value**	$Pr(>\|z\|)$	
Intercept	0.789	0.018	43.984	<0.001	***
Calmness (q1)	0.003	0.003	1.247	0.213	
Cheerfulness (q2)	-0.006	0.003	-1.938	0.053	
Concentration (q3)	0.005	0.002	2.338	0.020	*
Self-worth (q4)	0.010	0.004	2.779	0.006	**
Time of day [afternoon]	0.011	0.004	2.477	0.013	*
Time of day [morning]	0.003	0.004	0.643	0.520	
PHQ-9 depression score	0.004	0.001	3.105	0.003	**

* p <.05 ** p < .01 *** p < .001

Table C.2: Results of the linear mixed effect model using the *Mood questions*, the *Time of Day* and the *PHQ-9 score* to predict the average percentage of **Rules Completed** per session.

Model: *NumCorrectPct* $\sim$ *q1 + q2 + q3 + q4 + Time_of_day +* *(1|Subject) + (1|SessionID)*

Predictor	**Estimate**	**std. Error**	**t value**	$Pr(>\|z\|)$	
Intercept	0.795	0.013	61.206	<0.001	***
Calmness (q1)	-0.001	0.002	-0.388	0.698	
Cheerfulness (q2)	-0.004	0.002	-1.856	0.064	
Concentration (q3)	0.006	0.002	3.733	< 0.001	***
Self-worth (q4)	0.008	0.003	2.892	0.004	**
Time of day [afternoon]	0.013	0.003	4.208	<0.001	***
Time of day [morning]	0.023	0.003	7.354	<0.001	***

* p < .05 ** p < .01 *** p < .001

Table C.3: Results of the linear mixed effect model using the *Mood questions* and the *Time of Day* to predict the average percentage of **Tiles tapped correctly** per session.

Model: *RulesForgotten* $\sim$ *q1a + q1b + q2a + q2b + q3a + q3b + q4a + q4b +* Time_of_day + *(1|Subject) + (1|SessionID)*

Predictor	**Estimate**	**std. Error**	**t value**	$Pr(>\|z\|)$	
Intercept	3.596	0.998	3.605	<0.001	***
Calmness (q1a)	-0.001	0.096	-0.010	0.992	
Stressfulness (q1b)	0.024	0.089	0.266	0.791	
Cheerfulness (q2a)	-0.011	0.099	-0.116	0.908	
Sadness (q2b)	-0.080	0.103	-0.773	0.440	
Concentration (q3a)	0.190	0.095	2.006	0.045	*
Distraction (q3b)	0.332	0.088	3.791	<0.001	***
Self-worth (q4a)	-0.183	0.131	-1.402	0.161	
Depression (q4b)	-0.112	0.106	-1.055	0.292	
Time_of_day [afternoon]	0.016	0.165	0.100	0.921	
Time_of_day [morning]	-0.982	0.160	-6.133	<0.001	***

* p < .05 ** p < .01 *** p < .001

Table C.4: Results of the linear mixed effect model using the *Mood questions* and the *Time of Day* to predict the average number of **Rules forgotten** per session.

Model: *RulesRemembered* $\sim$ *q1 + q2 + q3 + q4 + Locale + PHQ_Score + (1|Subject) + (1|SessionID)*

Predictor	**Estimate**	**std. Error**	**t value**	$Pr(>\|z\|)$	
Intercept	3.688	0.407	9.072	<0.001	***
Calmness (q1)	0.105	0.061	1.713	0.087	
Cheerfulness (q2)	-0.123	0.065	-1.902	0.057	
Concentration (q3)	0.111	0.048	2.316	0.021	*
Self-worth (q4)	0.101	0.081	1.258	0.209	
Locale [en]	0.492	0.208	2.370	0.022	*
PHQ-9 depression score	0.080	0.031	2.629	0.011	*

* p < .05 ** p < .01 *** p < .001

Table C.5: Results of the linear mixed effect model using the *Mood questions*, the *Language used (locale)*, and the *PHQ-9 score* to predict the average number of **Rules remembered** per session.

Appendix D: Rules used in Wollie

Lvl	#	Rule description	Dutch translation
1	1	Tap numbers in descending order	Tik op getallen in aflopende volgorde
	2	Tap all things green	Tik op alle groene dingen
	3	Tap odd numbers	Tik op oneven getallen
	4	Tap nines	Tik op negens
	5	Tap animals	Tik op dieren
	6	Tap walruses	Tik op walrussen
	7	Tap monsters	Tik op monsters
	8	Tap green monsters	Tik op groene monsters
	9	Tap birds	Tik op vogels
	10	Tap tens	Tik op tienen
2	1	Tap numbers in ascending order	Tik op getallen in oplopende volgorde
	2	Tap twos	Tik op tweeën
	3	Tap ones	Tik op enen
	4	Tap bulls	Tik op stieren
	5	Tap birds	Tik op vogels
	6	Tap yellow birds	Tik op gele vogels
	7	Tap gingerbread men	Tik op peperkoekmannen
	8	Tap sixes	Tik op zessen
	9	Tap starfish	Tik op zeesterren
	10	Tap tens	Tik op tienen
3	1	Tap numbers in ascending order	Tik op getallen in oplopende volgorde
	2	Tap bunnies	Tik op konijntjes
	3	Tap green bunnies	Tik op groene konijntjes
	4	Tap odd numbers	Tik op oneven getallen
	5	Tap red dinosaurs	Tik op rode dinosaurussen
	6	Tap eggs	Tik op eieren
	7	Tap threes	Tik op drieën
	8	Tap sevens	Tik op zevens
	9	Tap blue starfish	Tik op blauwe zeesterren
	10	Tap monsters	Tik op monsters

Table D.1 continued from previous page

Lvl	#	Rule description	Dutch translation
4	1	Tap numbers in descending order	Tik op getallen in aflopende volgorde
	2	Tap lego blocks	Tik op legoblokjes
	3	Tap yellow butterflies	Tik op gele vlinders
	4	Tap fours	Tik op vieren
	5	Tap things you can drive in	Tik op dingen waar je in kunt rijden
	6	Tap green ducks	Tik op groene eenden
	7	Tap blue monsters	Tik op blauwe monsters
	8	Tap monsters with one eye	Tik op monsters met één oog
	9	Tap fires	Tik op vuur
	10	Tap wolves	Tik op wolven
5	1	Tap numbers in ascending order	Tik op getallen in oplopende volgorde
	2	Tap orange toys	Tik op oranje speelgoed
	3	Tap all rocking horses	Tik op alle hobbelpaarden
	4	Tap even numbers	Tik op even getallen
	5	Tap all things facing right	Tik op alle dingen die naar rechts wijzen
	6	Tap yellow TVs	Tik op gele tv's
	7	Tap fish	Tik op vissen
	8	Tap eights	Tik op achten
	9	Tap fires	Tik op vuur
	10	Tap yellow owls	Tik op gele uilen
6	1	Tap numbers in ascending order	Tik op getallen in oplopende volgorde
	2	Tap red bunnies	Tik op rode konijntjes
	3	Tap cars	Tik op auto's
	4	Tap sixes	Tik op zessen
	5	Tap yellow monsters	Tik op gele monsters
	6	Tap things that show teeth	Tik op dingen die tanden laten zien
	7	Tap light bulbs	Tik op gloeilampen
	8	Tap toy blocks	Tik op speelgoedblokken
	9	Tap green bulls	Tik op groene stieren
	10	Tap orange fish	Tik op oranje vissen

Table D.1 continued from previous page

Lvl	#	Rule description	Dutch translation
7	1	Tap numbers in descending order	Tik op getallen in aflopende volgorde
	2	Tap red butterflies	Tik op rode vlinders
	3	Tap all blue things	Tik op alle blauwe dingen
	4	Tap green monsters	Tik op groene monsters
	5	Tap all animals that can fly	Tik op alle dieren die kunnen vliegen
	6	Tap green toys	Tik op groen speelgoed
	7	Tap monsters with one eye	Tik op monsters met één oog
	8	Tap dinosaurs	Tik op dinosaurussen
	9	Tap yellow trucks	Tik op gele vrachtwagens
	10	Tap sixes	Tik op zessen
8	1	Tap numbers in ascending order	Tik op getallen in oplopende volgorde
	2	Tap all yellow things	Tik op alle gele dingen
	3	Tap red wolves	Tik op rode wolven
	4	Tap all things that are not animals	Tik op alle dingen die geen dieren zijn
	5	Tap green owls	Tik op groene uilen
	6	Tap all animals and monsters shown from the front	Tik op alle dieren en monsters getoond vanaf de voorkant
	7	Tap beavers	Tik op bevers
	8	Tap things that show teeth, which are not animals	Tik op dingen die tanden laten zien, maar die geen dieren zijn
	9	Tap red things in which you can drive	Tik op rode dingen waar je in kunt rijden
	10	Tap birds	Tik op vogels
9	1	Tap numbers in descending order	Tik op getallen in aflopende volgorde
	2	Tap all things facing left	Tik op alle dingen die naar links wijzen
	3	Tap green toys	Tik op groen speelgoed
	4	Tap all animals facing right	Tik op alle dieren die naar rechts wijzen
	5	Tap monsters	Tik op monsters
	6	Tap all animals that can fly	Tik op alle dieren die kunnen vliegen
	7	Tap all animals that howl	Tik op alle dieren die kunnen huilen
	8	Tap everything that is not an animal	Tik op alles dat niet een dier is
	9	Tap sevens	Tik op zevens
	10	Tap things with teeth showing	Tik op dingen die tanden laten zien

Table D.1: All the rules used in the nine levels of Wollie.

Appendix E: Images used in Wollie

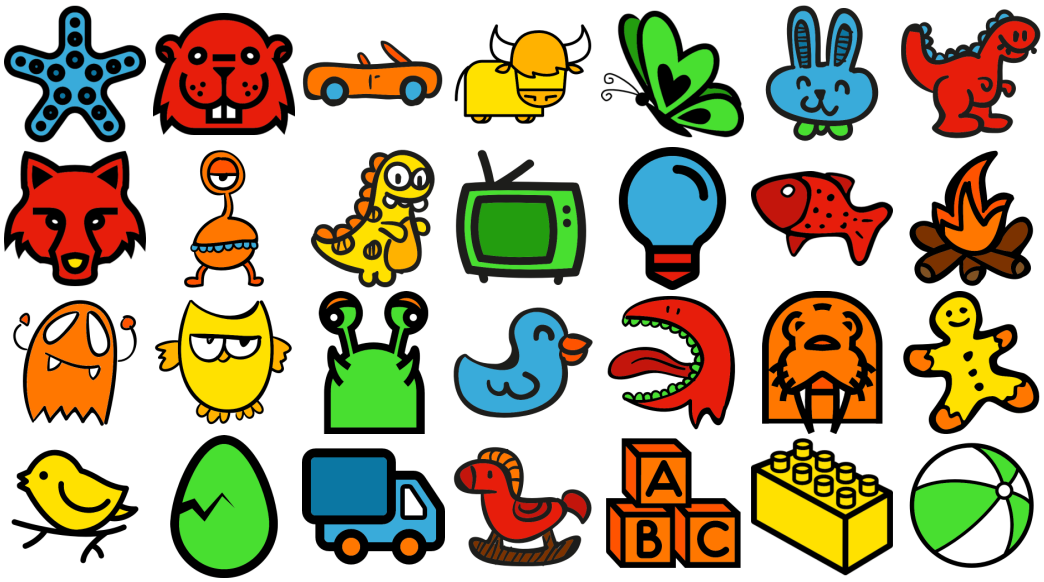
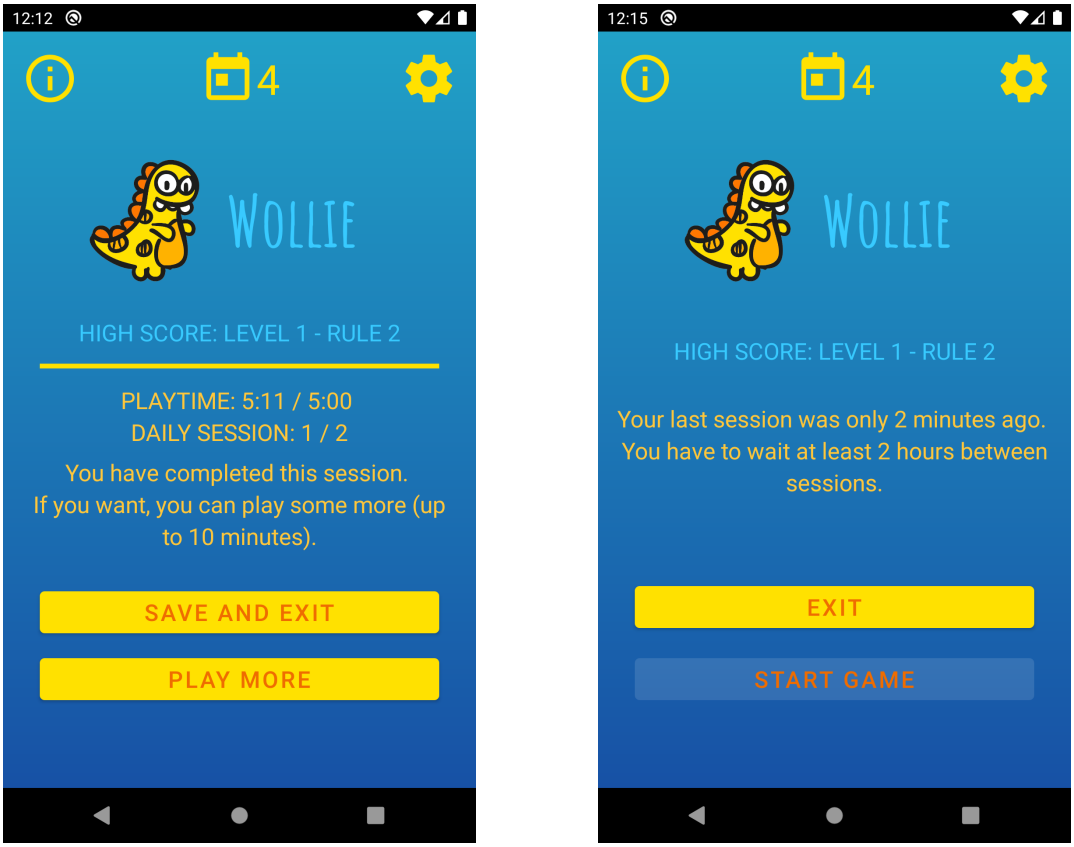


Figure E.1: All 28 images used in the smartphone game Wollie. Except for the fire image, every image has five color variations: blue, red, orange, yellow and green.

Appendix F: List of modifications for Wollie

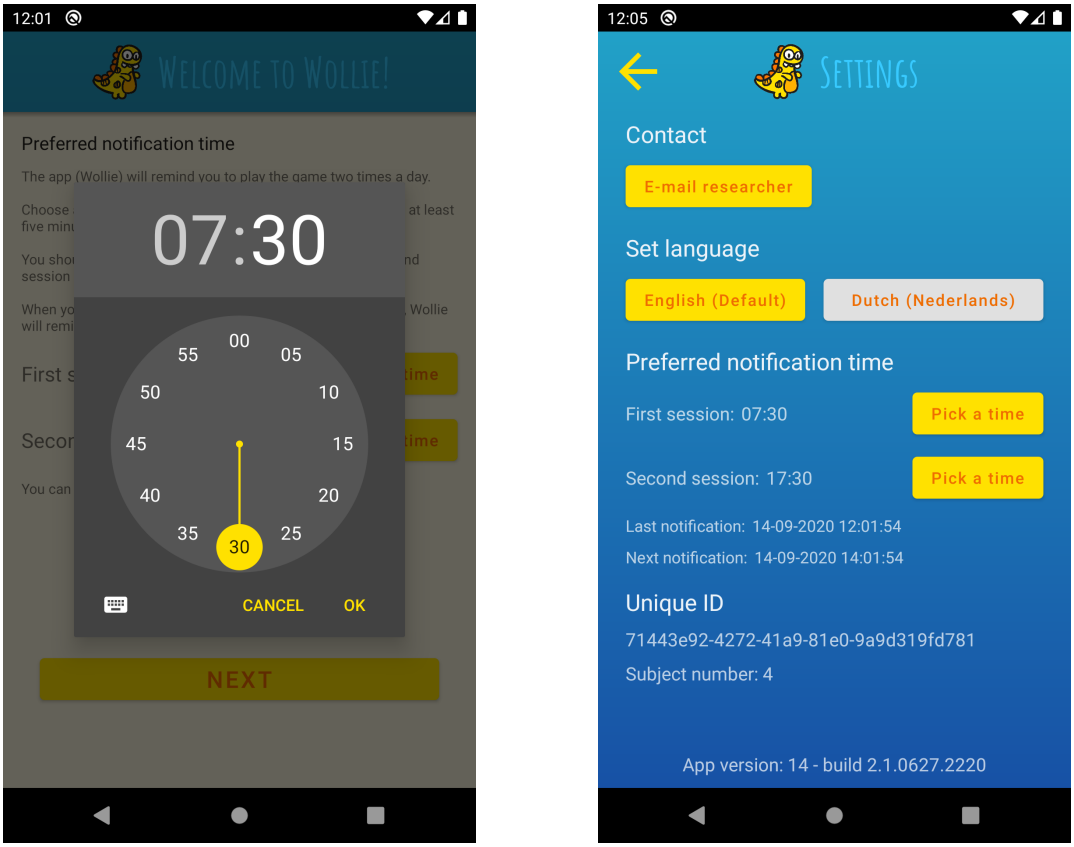
The original Wollie app was developed in 2015 (Doesburg & Taatgen, n.d.). A lot of changes were made to make the app suitable for the current study. Wollie had to be working on a wide range of Android smartphones with different screen sizes, including both older and newer Android versions. The most important changes are listed here:

- Added a short mood report questionnaire
- Allowed two playable sessions a day (between 7:00 and 23:45)
- Randomized the order of the rules in each level
- Dual language support (Dutch and English): everything in the app is translated
- Support for Android version 5.0 (Lollipop) and newer (up to Android 10 at the time of writing)
- Integration with Google FireBase to upload the anonymous log files to a secure cloud environment
- Push notification support with reminders to play the game at a custom time chosen by the user (see Figure F.2a)
- Settings screen added with options to contact the experimenter, check the app version, subject number, and change the language (see Figure F.2b).
- Added sounds and animations to make the app more modern and appealing to the user



(a) Session complete message on home screen **(b)** Two hour break message between sessions

Figure F.1: Screenshots of the home screen of Wollie.



(a) Preferred notification time screen **(b)** Settings screen

Figure F.2: Screenshots of some extra functions of Wollie.